Patterns

CellPress
OPEN ACCESS

Perspective

A guide to backward paper writing for the data sciences

Jon Zelner,[1,2,*] Kelly Broen,[1,2] and Ella August[1,3]
[1]Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA
[2]Center for Social Epidemiology and Population Health, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA
[3]Pre-Publication Support Service (PREPSS), Ann Arbor, MI 48109, USA
*Correspondence: jzelner@umich.edu

https://doi.org/10.1016/j.patter.2021.100423

THE BIGGER PICTURE As the size and complexity of datasets and data science pipelines continue to grow, the challenge of crafting clear and engaging scientific writing to communicate data science research has also increased. This comes at a time when effective communication of complex models and results is increasingly critical: from climate science to epidemiological modeling to the social sciences and beyond, data science involving very large datasets and complex, computationally intensive modeling is increasingly at the heart of the global scientific, policy, and public conversation. This paper is aimed at helping data science researchers successfully craft original research papers describing their work.

1 2 **3** 4 5 **Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

In this perspective, we outline a set of best practices for the planning, writing, and revision of scientific papers and other forms of professional communication in the data sciences. We propose a backward approach that begins with clearly identifying the scientific and professional goals motivating the work, followed by a purposeful mapping from those goals to each section of a paper. This approach is motivated by the conviction that manuscript writing can be more effective, efficient, creative, and even enjoyable—particularly for early-career researchers—when the overarching goals of the paper and its individual components are clearly mapped out.

INTRODUCTION

Academic and applied research in data-intensive fields requires the development of a diverse skillset, of which clear writing and communication are among the most important.[1] However, in our experience, the art of scientific communication is often ignored in the formal training of data scientists in diverse fields spanning the life, social, physical, mathematical, and medical sciences. Instead, clear communication is assumed to be learned via osmosis, through the practice of reading the work of others, writing up our own work, and receiving feedback from mentors and colleagues. What makes this way of learning frustrating for many researchers is that research papers in the quantitative sciences often have a relatively rigid format, and typically must be concise (i.e., 3,500 words or fewer), leaving little room for improvisation. In this perspective, we—a mid-career data science researcher in public health, advanced PhD student in quantitative epidemiology, and specialist in scientific writing—attempt to demystify the paper-writing process by providing a set of guiding questions that can be used to plan and revise each section of a paper and to help you position your paper to support your professional growth and goals.

Early-career data science researchers and practitioners may find themselves wondering how to work within the constraints of the research paper format to get their message across. In the worst case, the format is deadening, rather than freeing, and leads to papers that are as excruciating to read as they were to write. This problem becomes acute when the goal of publishing as communication is crowded out by the pressures of publishing for professional survival and advancement. When the volume and prestige of publications are prized over their value as tools of scientific communication—as they too often are[2,3]—the rigid format of the research paper can be just one more demotivating obstacle on a long and stressful career path. When the constraints and challenges of academic publishing are foregrounded, the process of writing can take on a sense of existential dread, with each unwritten manuscript section representing a new way to fail to achieve the standard needed to be successful.

Writing for data science poses unique challenges

Data science is an integrative enterprise that brings the processes of data cleaning, manipulation, visualization, and other forms of processing under the same scientific tent as statistical analysis and mathematical modeling.[4] Data scientists are often working at scale with "big data" for which the processes of procurement, storage, manipulation, and analysis is more complex than for smaller, less-complex datasets. This necessarily broadens the set of responsibilities the author of a data science paper has to explain a wide array of techniques, outputs, and results. To some extent, this mirrors age-old challenges in scientific writing: physical and natural scientists have long had to document their experimental setups, and social scientists present in-depth information on survey sampling and the collection of observational data. In fact, these fields have each become increasingly inflected by the challenges of working with large, complex datasets. So, while this guide is written with a data science audience in mind, we believe that the work of quantitative researchers in the sciences is increasingly that of disciplinary data science. We hope that the suggestions we provide will be applicable to the challenges of communicating increasingly complex analyses to as broad a scientific audience as possible.

What follows is a set of suggestions and questions to guide the process of writing up research for the data sciences. It is motivated by the strongly held conviction that the writing process should occur in parallel with the research process and that, in fact, the two are indistinct.[5] We also provide as an appendix an annotated version of a recent manuscript from our own research using Bayesian hierarchical models with complex public health datasets.[6] We use the annotation to show how the influence of these principles is reflected in a finished product. Our writing and professional advice is influenced by countless conversations and interactions with any number of colleagues, classmates, and mentors. While there are too many influences to thank, we are grateful for their cumulative wisdom, presented here in condensed form. We include citations to relevant resources and influences where possible (i.e., when the idea is presented in a paper, book, or blog post).

Each section of a research paper has a well-defined purpose

The relatively rigid format of research papers is often a pain point in scientific writing: What goes in the introduction versus the discussion? How much detail should be in the methods section? When should I mention limitations of my analysis and why? How do I do this without it coming off as stiff and formulaic? After a while, the inflexibility of the form may reveal itself to be freeing, because it provides a structure you can use to ensure that your ideas are clearly organized and communicated in a way that ensures that as much of your intended audience as possible can read, build on, and replicate.[7] The key thing to remember is that the sections of a research paper each have a distinct role to play in constructing and communicating your message, but they should also cohere and interact.[8,9]

Your paper is an extended argument about the relationship between a question (introduction), the way you decided to answer it (methods), what came out of it (results), what it all means (discussion), and what it implies for what is next (conclusion). If you start with a clear idea of what you want to accomplish with each of these sections, it will free you up to focus on your hypotheses, results, big ideas, and opinions.

Work backward to make sure your intended message gets across

A well-crafted data science paper is a pedagogical tool that not only conveys information from author to reader but facilitates the understanding of complex concepts. This works in both directions: The paper-writing process is an opportunity for the writers to learn about and clarify their understanding of the topic in addition to communicating it to someone else. If we can accept the idea of this kind of writing as teaching, we can take a lesson from research and practice in the field of educational development, particularly the backward approach to curriculum design, introduced by Williams and McTighe in their book *Understanding by Design*[10]:

> Our lessons, units, and courses should be logically inferred from the [learning outcomes] sought, not derived from the methods, books, and activities with which we are most comfortable. Curriculum should lay out the most effective ways of achieving specific results … the best designs derive backward from the learnings sought.

Under a backward-design approach, the overarching goals of a course are defined first, and then used to motivate and shape everything from the assignments students will complete, the nature and volume of reading material, and the way class meetings will be used to advance toward these goals. In this way of thinking, a course has a set of standard components—assignments, reading, class time—but the way in which they are devised and arranged is organized around supporting the learning goals of the class. The same approach can be applied to the construction of a research paper: even though most papers have the same sections (introduction, methods, results, discussion) early-career researchers may underestimate the amount of flexibility and room for creativity they have in using these components to achieve their scientific and professional development goals. The backward approach we lay out here is about starting at the end by answering the questions of "What do I want accomplish with this paper?" and scaffolding each piece to help serve those goals. This is contrasted with the more *ad hoc* forward approach most of us have learned to live with, in which we begin with the introduction and struggle through to the conclusion with the primary goal of simply finishing the manuscript.

A GUIDE TO BACKWARD PAPER WRITING

While writing this guide, we struggled with what to call it: a checklist implies something proscriptive, a set of rigid "must-do" tasks, much as a pilot has a pre-flight checklist where each item is essential to a safe journey. However, the goal of writing for clear and impactful scientific communication is not served by ticking off externally imposed requirements without knowing why one is doing so.[11] Good scientific writing should still tell a story that brings the reader along on a condensed version of your journey with a given project: why is the idea compelling and important? Why did you do what you did? What do you wish you could have done? What should you or someone else do next?[1]

Unfortunately, the hierarchical nature of many academic and non-academic research environments can result in scientific writers at all stages fixating on the voice of an internalized critic: the unsupportive professor, supervisor, or colleague who aggressively brushes you back if you get too confident. This can lead to defensive, apologetic writing through which we are pursuing the goal of not being criticized rather than the goals that brought us to the work in the first place. Our overarching objective in this piece is to help you keep focus on making an honest, affirmative, and enthusiastic argument in favor of the work you have done. To facilitate this, in what follows, we have constructed a set of questions to ask yourself and your coauthors throughout the process of planning, writing, submitting, and revising your paper. However, like any good guiding questions, they are only useful if you engage thoughtfully with them until they are answered to your satisfaction.

Questions to help you plan your writing, reflect on your professional goals with the paper, and position your work

Before you get started, take some time to be sure you have a good sense of your answers to the following questions. Often, it can be helpful to make a fresh document in which you think out these sorts of high-level questions on the page. This sort of informal ‘‘pre-writing’’ is powerful because it can help you clarify your thoughts and professional goals[12]:

1. **Who is your intended audience?** This could also be described as, "What do you want to accomplish with your paper?" If you have a clear idea of who you hope will read your paper, it will help clarify the goals of your writing[13]: is the goal to showcase incremental improvement to existing approaches to a group of people who all specialize in the same area? Are you trying to reach a broad audience to convince them why some long-held idea is not quite right? Do you want to reach a specialist and non-specialist audience to contribute to policy or public conversation?
2. **What is the major idea/contribution of the paper?** If you have more than one, you either need to reconsider what your main idea is or think about splitting into two papers. The reason to do this is simple: you have not much space and not much attention from other people and you want to make the most of what you have. Sticking to one idea per paper[14] is, in general, a good practice to ensure that your message comes through as clearly and impactfully as possible.
3. **What are your professional goals in writing this paper?** Your career stage and level of engagement with the topic can and should dictate the way you approach a paper. Is this the publication you want to showcase when looking for your next job? Is this part of building a research program on your own by testing out new ideas?
4. **What is the right outlet for this work?** Each journal has a core readership, as well as style, organization, and length requirements. Two important questions that should guide your choice of target journal are who the best audience is for your paper, and how much value you place on publishing in a prestigious journal (and your ability to tolerate the time investment and higher risk of rejection associated with aiming high) versus starting the process with an outlet that is less of a ‘‘reach’’ but can still let you reach the audience you want to connect with.[8] Identifying the core readership of a journal, which is often laid out in bold print on the journal’s website, will help you reach those who are interested in your work and those who are poised to apply your findings.

Below, we describe each section of an academic journal article. We present these in the order they typically show up in a published paper, but we do not have hard-and-fast rules about the order in which they should be written: some writers prefer to start with the introduction section, and others prefer to start with the methods section. Experiment to find out what feels right for you.

INTRODUCTION

5. **What is the problem your analysis is meant to address or solve?** Use the first paragraph or two to outline the scope and importance of the problem you are addressing. The importance may encompass the burden of disease (or other outcome), such as suffering and or mortality, and/or economic impact, as well as the scope or scale of the disease. The problem you articulate should be the specific knowledge gap that you will fill with your research, and the reader should come away with a good sense of why it is important to close it.
6. **What else has been tried?** This is the part where you get into the approaches that have been taken to this problem and the results that have come out of this.
   a. *No need to be negative*: This is not about why everything else is awful, but just what has already been accomplished, what others have taught us with their previous work. If you have prior work in this area, this is also a good place to cite it and highlight the continuity between your earlier research and what you are doing now.
7. **Nevertheless**… This is where you can identify the gap left by the previous work. For example, it may fit into one of these example categories:
   a. Previous analyses got at an *important question* but not with the kind of data you are using, which may be more detailed, contextually relevant, etc.
   b. Previous work used a *methodological approach* that was not able to get at some important dimension of the problem, over/under-stated variability in outcomes, etc.
   c. Earlier work did not address the *broader context* of the problem; i.e., too narrowly focused on a specific dataset or place and less on understanding the processes that cut across contexts.
8. **In this paper we will**… For the love of everything, please tell the reader what you are going to do before you do it. This is the point of departure for your reader on the journey that will be reading and metabolizing your paper: give them a map!
   a. *Use this part to mention the data you will use*; i.e. where it comes from and what outcomes measured by the data are the focus of your analysis.

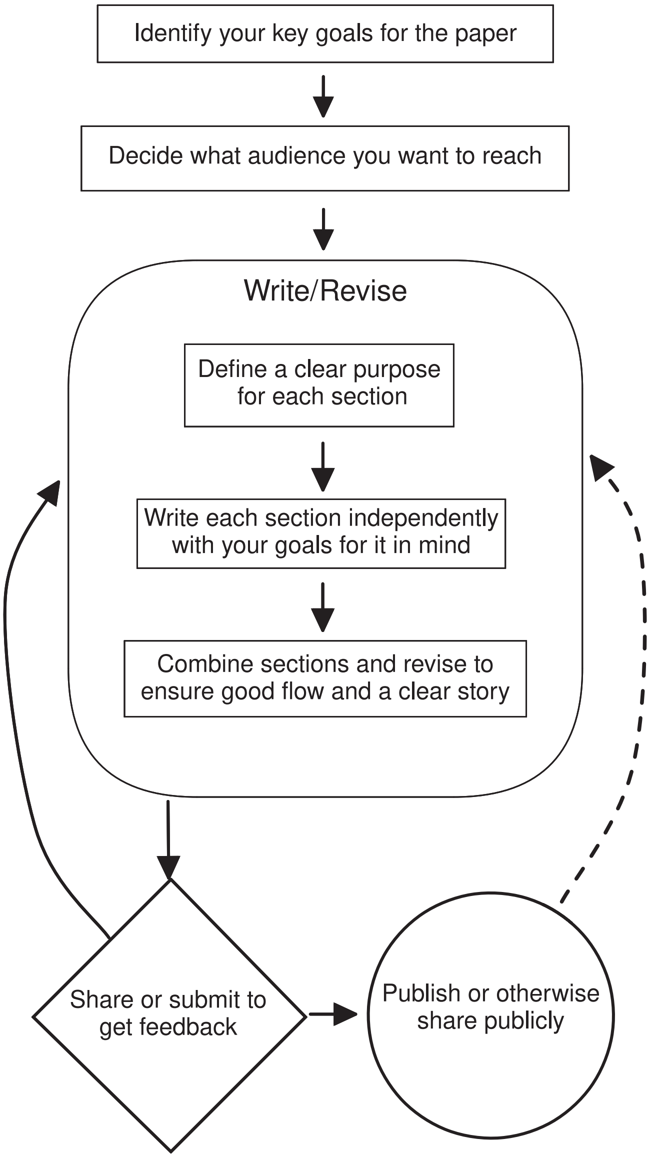
Identify your key goals for the paper
Decide what audience you want to reach
Write/Revise
Define a clear purpose for each section
Write each section independently with your goals for it in mind
Combine sections and revise to ensure good flow and a clear story
Share or submit to get feedback
Publish or otherwise share publicly

Figure 1. Schematic representation of the process of backward paper writing

The high-level steps involved in the process of backward data science manuscript preparation. The square boxes at the top represent the important pre-writing steps in which you clarify the scientific and professional goals motivating your work. The rounded box represents the process of initial writing and revision. Once a draft is complete, the diamond box represents circulating the manuscript to colleagues and mentors for feedback, or submitting for publication, with the expectation that this will result in further revision and updating of your work. The circle represents the typical endpoint of the process: publication in a peer-reviewed outlet, sharing publicly via a preprint server, publishing online via an interactive notebook or app, or the many other ways in which data science research can be disseminated to relevant scientific communities and the public at large. Finally, the dashed arrow represents the potential for post-publication revision in response to feedback and critique or new data. While not required, this type of post-publication revision is increasingly common in data science fields, allows for greater transparency, and may increase the long-term relevance of the published work.

b. *Describe the analytic/statistical methods in brief*. Do not get into a ton of depth, just mention what you are going to do and provide a one-sentence justification. This allows the reader to have a sense of what is coming in terms of the analytic approach without getting into information that should be in the methods section. This kind of signposting is often very valuable for giving your reader a sense that they know where they are going when reading your paper rather than feeling lost or confused.

METHODS/DATA

The methods and data section can be tricky to write because it is where all the things you are trying to do converge, and where an interested reader or reviewer will spend a lot of their time trying to understand what you have done and verifying that the results and conclusions are justified. So, in this section we are a bit more detailed and proscriptive than in the others, since there is so much packed into a relatively small section of your paper:

1. **Introduce the methods and data in broad terms.** Use the first paragraph of the methods and data section to do some signposting; i.e., giving the reader a sense of the reasoning behind the study design, and motivation for using a particular dataset or focusing on a particular population to answer your particular question.
   a. *This applies even to methods papers:* In most cases, the introduction to the entire paper should be focused primarily on the applied problem at hand. If it is a more methods-focused or statistical paper, that may not entirely be the case. But even then, the opening of the methods section should focus on motivating the importance of the method for confronting real-world problems, rather than getting into the nuts and bolts of the approach.
2. **Go into detail on the data.** This section can and should be a bit more formulaic than the ones that came before. It really is as much "just the facts ma'am" as possible, but the trick is highlighting the facts about the data that are most relevant to what you are trying to do. The motivating question in putting this section together—which can include figures, tables, and written description—should be, "What does the reader need to know about the data to understand the results?"
   a. *What do the data measure?* Include outcomes, covariates, etc. If you have multiple sources (e.g. cases of a disease plus census population data), introduce both.
   b. *When/how were they collected? Who collected the data?* This is what is sounds like: give a sense of the provenance of the data. If they come from a larger/long-term study, refer to earlier analyses from the study. If a published study protocol exists, be sure to cite it. If the data are publicly available, link to them and be sure to include a timestamp showing when you retrieved them.
3. **What can go into a figure?** Visualizing data is almost always preferable to talking about them or presenting summary statistics in a table,[15] but it requires care: think about what views of the data are most important for a descriptive figure that introduces the data and the problem (Figure 1). Remember that the point is *not* to make the reader as much of an expert about the data as you are, but to be familiar enough with the data to understand both the motivation for the analysis and the results when they come along. (See Nolan and Stoudt, 2021[1] for a more in-depth exploration of the figure-making process.)
4. **Do I need a table?** Really, do you need a table?[16] If you do, keep it short and sweet: present relevant information that is hard to get into a single figure (e.g., cramming the number of people in a study, the proportion in different age groups, and the distribution of individuals by sex, race/ethnicity, etc. into a figure may be needlessly complex and better accomplished in a table).

Analytic methods

As methods-oriented people, we often want to get into a lot of detail here (we will speak for ourselves at least!). However, this is another moment where you need to consider your audience and goals carefully. You are probably doing a bunch of things in the paper that merit explanation in the methods section, but the ones you highlight in this section should fit the following criteria:

9. **What is the most important thing you are trying to do with your paper, and which methods are most crucial to understanding that?** This is another way of saying that you should focus on the model or

set of models in the paper that are most important to understanding your results and advancing your overall agenda with the paper. You may have sub-analyses that provide extra detail but do not need to be outlined in depth in the main part of the manuscript. If it cannot all fit in the main text without it getting bloated, you can always refer the reader to a supplement containing this information.

10. **Which methods are important enough to describe is a function of your audience.** Think about what kind of paper this is and who you expect to read it; that will dictate what goes in this section. The more generic the audience is, the more likely it is that you will have a lot of the methods in a supplement, whereas, for a more specialist audience, you will probably go into more depth. For example, if your analysis uses a fancy new statistical technique and your goal is to highlight what more can be learned using this approach to other people working in the same area, then you might want to go into depth. However, if the point is to highlight the scientific importance of the result yielded by the method for a broader, non-specialist audience, you may save more of that detail for a supplement.
11. **How does what you are doing methodologically relate to the major questions of your analysis?** Remember that the methods section exists to give the reader the ability to understand and evaluate the results you are presenting to them. So, if the result of using a particular method is not going to be in the results in some way, it should not be in the methods section.

RESULTS

The point of the results section is to make the results of your analysis (descriptive statistics, assays, imaging results, parameter estimates, posterior predictions, model simulations, etc.) as clear as possible to the reader. The first part of that is figuring out the best way to communicate each relevant piece of information. We would boil this down into a set of simple suggestions:

12. **If it can be conveyed visually, do it!** Prefer figures over tables and in-text descriptions where you can.[15,17] This is subject to limitations that force you to prioritize what goes into a figure and what does not: How many figures are allowed by the journal? Is it enough information to take up a whole figure or would briefly mentioning it in the text be a better use of space?
    a. *Figures and tables should stand on their own.* Reasonably informed readers should be able to get what is going on from looking at your figure and reading the legend, even if they have not read the rest of the paper. This is not a hard-and-fast rule, but if you work toward it you will ensure that the figures convey as much information as possible.
    b. *Each figure should make a clear point of its own.* If two separate figures convey overlapping information, try to eliminate one, or combine them both into a single panel (e.g., using left- and right-hand axes) or a multi-panel figure. Each full figure (i.e., with its own number) should touch on a single idea/result. For example, if you are reporting the results of an analysis that looks at multiple outcomes (e.g., the risk of developing a disease versus the risk of dying of it), it likely makes sense to put these in separate figures, much as you would place the discussion of them in separate paragraphs in the text.
    c. *Use your figures and results to tell a story.* Nolan and Stoudt suggest using a process of “storyboarding” in which you arrange your figures and tables in the order you think they make the most sense, and then write up your results in a way that takes you through each to tell the story of your analysis.[1] Just like the process of storyboarding for TV and movies is an iterative one, the idea here is to give yourself maximum flexibility to re-arrange the pieces until you find a narrative that brings it all together.
13. **If you must make a results table, keep it small and simple.** Big, complex tables are where reader attention goes to die. If information is best conveyed by a table, be sure to include only the most essential information. When a table gets too big, it becomes easy to forget what its purpose is. By keeping it short and cutting out extraneous information, you are better able to keep the focus on your message.
14. **Use the beginning of the results section to hit the highlights in the figures and tables.** Imagine you are explaining the figures to someone: what is the most important thing you want them to get from the figure? Talk about that in the beginning of the results section. Whatever you do, do not recapitulate entire figures and tables. They are part of your results, you can and should refer to them, but they should be complementary to what you are writing here, not duplicative or completely disjointed from that.
15. **Use the remaining text of the results section to provide information *not* in the figures or tables.** What else is important to know that is not captured by a figure or table? Is there a single estimate from a side analysis that fills in the story but does not warrant a figure or table on its own?

Ethical reporting of methods and results

Often, when we investigate a question, we attempt multiple analytic approaches before deciding on the one presented in the final paper. This is in the nature of the research process and is not inherently problematic. However, it can present a conundrum when the time comes to write up your paper: the short format and the need to tell a compelling story can result in papers that make a complex data science project appear to have occurred in one straight line from hypothesis to data to methods to results. It is important to think carefully about which of these detours is important enough to include in the main text versus the supplementary materials or some other product associated with your paper.[18]

You can also use the data science ecosystem to your benefit. Rather than fighting the constraints of the research paper structure, it can help to think of your paper as a single node in a larger network of outputs that provide transparency and reproducibility. For example, by releasing the code to complete your analysis, and the data underlying it if it is possible to do so without sacrificing privacy,[19] you give other researchers the ability to assess your research products directly by re-running the code used to generate your results. However, you could also describe your process in greater depth in a blog post or a standalone essay submitted to a preprint server like Arxiv.org. You could release a step-by-step walk-through of your analysis using Jupyter notebooks, Shiny apps for R, or any other tool that facilitates more interactivity than a static PDF.

This kind of transparency allows you to illustrate and defend your approach and to also explain how you accounted for prior explorations in your presentation of results. This can help you increase confidence that your major results are not a reflection of data-driven hypothesis-testing, i.e., p-hacking or a trip through the "garden of forking paths," in which researchers make many comparisons before deciding on the ones to present in a manuscript.[20]

DISCUSSION

The discussion is where you get to be a bit more expansive and opinionated. As ever, though, think about how each of the things you put in here will affect what you are trying to accomplish. We tend to think of the discussion as having its own subsections[7,9] that roughly look like this:

16. **First paragraph of discussion: Summarize, summarize, summarize.** What did you accomplish? How did your results relate to the problem/hypotheses you laid out in the introduction? If I had not read the introduction, methods, or results, and just opened it to the discussion and read this paragraph, I should be able to get what you did.
17. **Second paragraph of discussion: Sell the product.** This is it: time to make the affirmative case for what you did. Why is it important? Why was your approach well suited to answering the question? What gaps have you worked toward closing that you highlighted in the introduction? Again, there is no need to be negative about other work, just show how you have moved the research forward in some meaningful way.
18. **Third paragraph: Limitations.** Writing this part can be uncomfortable or scary sometimes because it seems like you are being asked to undermine your work. However, a well-written limitations paragraph adds to your credibility by showing you have thought about what can and cannot be learned using the data available and the methods employed. Try to answer the question of, "What question might someone else want to answer that my paper does not/cannot address?" Limitations are reasonable stopping points for what you are trying to do that demarcate the boundaries of your analysis. Sometimes a question that should have been asked was not included in a survey, or an instrument malfunctioned during an experiment that was too costly or time consuming to be repeated. Being forthright about these limitations is also a helpful reminder to your readers (and yourself) that you cannot solve every problem, and that your results reflect the limits of the tools and data available to you.
    a. *Future directions.* Once you have given the reader a sense of what your analysis does and does not do, you can talk about how you, or someone else, can take the next step to transcend the limitations you have highlighted. This is an opportunity to plant your flag on the things you want to do next or that you have just started doing. You can use this as a form of motivation and accountability: What did I not do here that I want to do next?
19. **Fourth paragraph: Conclusions** (sometimes part of the Discussion, sometimes a separate section of the paper). This is where you tie everything together: What does it all mean and why does it matter? What comes next? What is the big point you want to stick with the reader? Here is where you can get a bit more opinionated, editorialize a bit, and even make some tentative extrapolations or predictions about the future.

Accept—and anticipate—the process of revising your work

The process of submission, rejection, and resubmission is often deeply discouraging to early-career researchers. A common misconception is that the best papers sail through review without revision or critique from reviewers, editors, and colleagues. However, it is in fact this process of feedback and revision that often makes good papers great.[21] It cannot be reiterated enough that critiques or suggestions for improvement to your work do not imply that you have failed and that it is unpublishable.[22,23] We hope that following a process like the one outlined here, whether you are an early-, mid- or late-career researcher, will allow you to remain motivated throughout what can be a stressful and oftentimes perplexing process.

Make it your own

We hope that the outline provided here is helpful and provides a roadmap, or at least a set of guardrails, for your writing. It is important to remember that even scientific writing is a personal and creative process. Our initial goal in writing this guide was to make that process easier, less painful, and even a bit fun for ourselves. To facilitate the adoption and customization of the approach outlined here by data science researchers across the scientific spectrum, several versions of these questions are available at https://github.com/epibayes/paper-template. In addition to providing a version in Word, versions in Markdown and RMarkdown are available as well. This way, you can download a copy and begin filling in the sections as you go, using the questions as a scaffold for your research and writing. You are also invited to edit and remix this guide based on your own experiences and interests, with the hope that you will share your insights with colleagues and the next generation of researchers.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100423.

AUTHOR CONTRIBUTIONS

J.Z., K.B., and E.A. all contributed to the conceptualization, writing, and revision of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Nolan, D., and Stoudt, S. (2021). Communicating with Data: The Art of Writing for Data Science, 1st edition (Oxford University Press).

2. Curry, S. (2018). Let's move beyond the rhetoric: it's time to change how we judge research. Nature *554*, 147. https://doi.org/10.1038/d41586-018-01642-w.

3. Stern, B.M., and O'Shea, E.K. (2019). A proposal for the future of scientific publishing in the life sciences. PLoS Biol. *17*, e3000116. https://doi.org/10.1371/journal.pbio.3000116.

4. van der Aalst, W. (2016). Data science in action. In Process Mining: Data Science in Action, W. van der Aalst, ed. (Springer), pp. 3–23. https://doi.org/10.1007/978-3-662-49851-4_1.

5. Schulte, B.A. (2003). Scientific writing & the scientific method: parallel “Hourglass” structure in form & content. Am. Biol. Teach. *65*, 591–594. https://doi.org/10.2307/4451568.

6. Zelner, J., Trangucci, R., Naraharisetti, R., Cao, A., Malosh, R., Broen, K., Masters, N., and Delamater, P. (2020). Racial disparities in COVID-19 mortality are driven by unequal infection risks. Clin. Infect. Dis. *72*. https://doi.org/10.1093/cid/ciaa1723.

7. Iskander, J.K., Wolicki, S.B., Leeb, R.T., and Siegel, P.Z. (2018). Successful scientific writing and publishing: a step-by-step approach. Prev. Chronic Dis. *15*, 180085. https://doi.org/10.5888/pcd15.180085.

8. Busse, C., and August, E. (2020). How to write and publish a research paper for a peer-reviewed journal. J. Canc. Educ. *36*, 909–913. https://doi.org/10.1007/s13187-020-01751-z.

9. Lang, T.A. (2017). Writing a better research article. J. Public Health Emerg. *1*, 88. https://doi.org/10.21037/jphe.2017.11.06.

10. Wiggins, G., and McTighe, J. (2005). Understanding by Design, 2nd Expanded edition (Assn. for Supervision & Curriculum Development).

11. Leape, L.L. (2014). The checklist conundrum. N. Engl. J. Med. *370*, 1063–1064. https://doi.org/10.1056/NEJMe1315851.

12. Estrim, H. (2015). Disciplinary and professional identities are constructed through writing. In Naming what We Know: Threshold Concepts of Writing Studies, E. Wardle and L. Adler-Kassner, eds. (University Press of Colorado), pp. 55–57. https://muse.jhu.edu/book/40635.

13. Ede, L., and Lunsford, A. (1984). Audience addressed/audience invoked: the role of audience in composition theory and pedagogy. Coll. Compost. Commun. *35*, 155–171. http://www.jstor.org/stable/358093.

14. Lenz, M. (2018). One idea per paper! Handling ideas. https://handlingideas.blog/2018/09/01/one-idea-per-paper/.

15. Tufte, R.E. (2001). The Visual Display of Quantitative Information, 2nd edition (Graphics Press).

16. Figures and Charts. The Writing Center ● University of North Carolina at Chapel Hill. https://writingcenter.unc.edu/tips-and-tools/figures-and-charts/.

17. Gelman, A. (2020). Regression and Other Stories, 1st edition (Cambridge University Press).

18. American Statistical Association (2018). Ethical guidelines for statistical practice. https://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf.

19. Broen, K., Trangucci, R., and Zelner, J. (2021). Measuring the impact of spatial perturbations on the relationship between data privacy and validity of descriptive statistics. Int. J. Health Geogr. *20*, 3. https://doi.org/10.1186/s12942-020-00256-8.

20. Gelman, A., and Loken, E. (2014). The statistical crisis in science. Am. Sci. *102*, 460. https://doi.org/10.1511/2014.111.460.

21. Downs, D. (2015). Revision is central to developing writing. In Naming what We Know: Threshold Concepts of Writing Studies, E. Wardle and L. Adler-Kassner, eds. (University Press of Colorado), pp. 66–67. https://muse.jhu.edu/book/40635.

22. Taczak, T. (2015). Reflection is critical for writers’ development. In Naming what We Know: Threshold Concepts of Writing Studies, E. Wardle and L. Adler-Kassner, eds. (University Press of Colorado), pp. 78–83. https://muse.jhu.edu/book/40635.

23. Bean, J.C., and Melzer, D. (2021). Engaging Ideas: The Professor’s Guide to Integrating Writing, Critical Thinking, and Active Learning in the Classroom, 3rd edition (Jossey-Bass).

About the authors

Jon Zelner is an infectious disease epidemiologist and assistant professor of epidemiology at the University of Michigan (UM) and leads the EpiBayes research group at the University of Michigan (epibayes.io).

Kelly Broen is a PhD student in epidemiology at the University of Michigan researching the spatial and social epidemiology of infectious disease.

Ella August is a specialist in scientific writing for the health sciences, clinical associate professor of epidemiology at the University of Michigan, and the editor-in-chief of the Pre-Publication Support Service (PREPSS; sites.google.com/umich.edu/prepss).

Patterns, Volume *3*

Supplemental information

A guide to backward paper writing for the data sciences

Jon Zelner, Kelly Broen, and Ella August

Paper written for audience concerned with understanding determinants of social inequalities in mortality from COVID-19, including researchers, policymakers & journalists. Clinical Infectious Diseases appropriate due to wide readership in public health and medicine and credibility as reliable outlet for high quality research to news media.

Clinical Infectious Diseases

MAJOR ARTICLE

Title clearly states most important result. Clear message essential to goal of correcting discourse on determinants of race/ethnic inequalities in COVID-19 mortality. Throughout most of 2020, there was misunderstanding that these differences came largely from different rates of case-fatality, i.e. risk of dying after infection. Our paper added to growing understanding that most of this inequity can be explained in terms of exposure, which can be addressed through policies that give people equal protection against infection regardless of race/class.

Racial Disparities in Coronavirus Disease 2019 (COVID-19) Mortality Are Driven by Unequal Infection Risks

Jon Zelner,[1,2] Rob Trangucci,[3] Ramya Naraharisetti,[1,2] Alex Cao,[4] Ryan Malosh,[1] Kelly Broen,[1,2] Nina Masters,[2] and Paul Delamater[5]

[1]Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan, USA, [2]Center for Social Epidemiology and Public Health, University of Michigan School of Public Health, Ann Arbor, Michigan, USA, [3]Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA, [4]Consulting for Statistics, Computing and Analytics Research, University of Michigan, Ann Arbor, Michigan, USA, and [5]Department of Geography, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Background. As of 1 November 2020, there have been >230 000 deaths and 9 million confirmed and probable cases attributable to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the United States. However, this overwhelming toll has not been distributed equally, with geographic, race/ethnic, age, and socioeconomic disparities in exposure and mortality defining features of the US coronavirus disease 2019 (COVID-19) epidemic.

Methods. We used individual-level COVID-19 incidence and mortality data from the state of Michigan to estimate age-specific incidence and mortality rates by race/ethnic group. Data were analyzed using hierarchical Bayesian regression models, and model results were validated using posterior predictive checks.

Results. In crude and age-standardized analyses we found rates of incidence and mortality more than twice as high than for Whites for all groups except Native Americans. Blacks experienced the greatest burden of confirmed and probable COVID-19 (age-standardized incidence, 1626/100 000 population) and mortality (age-standardized mortality rate, 244/100 000). These rates reflect large disparities, as Blacks experienced age-standardized incidence and mortality rates 5.5 (95% posterior credible interval [CrI], 5.4–5.6) and 6.7 (95% CrI, 6.4–7.1) times higher than Whites, respectively. We found that the bulk of the disparity in mortality between Blacks and Whites is driven by dramatically higher rates of COVID-19 infection across all age groups, particularly among older adults, rather than age-specific variation in case-fatality rates.

Conclusions. This work suggests that well-documented racial disparities in COVID-19 mortality in hard-hit settings, such as Michigan, are driven primarily by variation in household, community, and workplace exposure rather than case-fatality rates.

Keywords. COVID-19; SARS-CoV-2; social epidemiology; disparities.

As of 1 November 2020, there have been more than 230 000 deaths and 9 million confirmed and probable cases attributable to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the United States, with these numbers undoubtedly reflecting a substantial underestimate of the true toll. Geographic, race/ethnic, age, and socioeconomic disparities in mortality have been key features of the first, second, and ongoing third wave of the US coronavirus disease 2019 (COVID-19) epidemic [1–5]. However, the extent to which this differential mortality is driven by disparities in rates of infection by age, race, and socioeconomic status (SES), or some combination thereof, remains unknown. Addressing the clear inequities in the toll of death resulting from the COVID-19 pandemic in the United States requires disaggregating the relative role of exposure leading to infection from age-specific case-fatality rates in drivers of the gaping inequity characteristic of SARS-CoV-2 mortality in the United States.

Analyses of other respiratory viruses, such as respiratory syncytial virus and influenza, have documented race/ethnic disparities in both rates of infection and case fatality [6]. This inequality is driven by diverse factors including comorbid conditions that increase susceptibility to infection and disease severity. But it is also a function of structural factors that impact the ability of members of different race/ethnic and socioeconomic groups to avoid infection. Relevant factors include mass incarceration [7, 8], residential segregation [9, 10], and wealth inequality that facilitates social distancing among the well-off while poorer individuals are more likely to be compelled into "essential work" [11]. A recent cross-national systematic review placed the population average infection fatality ratio of COVID-19 infection at 0.75% [12]. However, demographic factors such as population age structure are key shapers of such rates and their variation across social groups [13]. While some studies have illustrated the differential impact of SARS-CoV-2 on non-White populations in the United States using aggregated data [5], no existing

Clearly identified problem we address with paper.

Received 14 September 2020; editorial decision 5 November 2020; published online 21 November 2020.

Correspondence: J. Zelner, University of Michigan School of Public Health, Department of Epidemiology, 1415 Washington Heights, Ann Arbor, MI 48109 (jzelner@umich.edu).

Clinical Infectious Diseases® 2021;72(5):e88–e95

DOI: 10.1093/cid/ciaa1723

“In this paper we will…” statement that identifies objectives. Readers deciding whether to read on may decide here if interested; important to be clear and concise.

analyses provide a clear breakdown of these risks by age, sex, and race [14]. In this paper, we aim to partially close this gap using detailed case-level data from the US state of Michigan, which was particularly hard hit by SARS-CoV-2 in the winter and spring of 2020, and where the epidemic has been marked by unmistakable racial and socioeconomic inequality.

Clearly identified problem we address with paper, cont.

METHODS

Data

Our data is fairly intuitive for intended audience; rather than providing deep description, we jump into nuts and bolts of our data.

We used data from 73 441 people with polymerase chain reaction (PCR)–confirmed and probable COVID-19 infections recorded by the Michigan Disease Surveillance System (MDSS) from 8 March 2020 through 5 July 2020. Probable cases were defined using the criteria outlined in the Michigan State and Local Public Health COVID-19 Standard Operating Procedures [15]. From this dataset, we excluded 25 cases who did not reside in Michigan or were missing a state of residence, 8613 people for whom race or ethnicity was not recorded, and 27 people who did not have age recorded or had an age more than 116 years old, indicating entry errors. We combined 68 pairs of records that had duplicate patient identification numbers, resulting in 34 fewer cases. Finally, we dropped 28 patients whose sex at birth was unknown, leading to a final dataset of 49 701 people with a confirmed or probable COVID-19 infection and with known age, race or ethnicity, state of residence, sex at birth, and state prisoner status. To mitigate the potential of right-censored deaths to erroneously deflate mortality rates, we truncated the data at the 97.5% quantile of the distribution of times to death from case referral date, which was 46 days, after which our data comprise 58 428 individuals.

Focus on data processing steps that are important to understanding outcomes. Definition of race/ethnicity and treatment of age groups here is critical for interpreting results, therefore we included this information in main text.

After filtering the case data, we binned age by 10-year intervals to age 80, with ages 80 and above in 1 bin. We also assigned cases to race/ethnicity categories of Black/African American, Latino, Asian/Pacific Islander, Native American, Other, and White, where Other comprised the census category of “other” and mixed-race individuals. To model per-capita rates of disease we used IPUMS public-use microdata from the 2018 American Community Survey (ACS) [16] to obtain population counts for each age/sex/race stratum. For additional information on data preparation and assignment of cases to race/ethnic categories, see Section 2.1 of the Supplementary Materials.

COVID-19 Cumulative Incidence Rates

To calculate age-specific, per-capita rates of COVID-19 infection in each age (i), sex (j), and race (k) bin, we fit a Poisson regression model with a population offset term, $\log(n_{ijk})$, where n_{ijk} is the size of the population for the ijk-th group from the 2018 ACS. We included age × sex, age × race, and sex × race interaction terms to capture the full spectrum of potential heterogeneity in our outcome data. We denote the observed number of cases in each group as y_{ijk} and the per-capita cumulative incidence rate in each bin as λ_{ijk}. To ensure comparability of incidence and mortality rates across race/ethnic groups, we employed a direct standardization approach to provide age- and sex-adjusted results where necessary. For all analyses of per-capita age-specific incidence rates, we used a log-Gaussian prior distribution with a mean of 0 and standard deviation of 0.1.

Gave minimum amount of information a reader with general understanding of analytic methods in public health and epidemiology would need to interpret results.

Analysis covers two outcomes: 1) infection with SARS-CoV-2 (virus that causes COVID-19); 2) death from COVID-19. We clearly identified outcomes with subheadings to highlight that different regression models were applied to different outcomes.

Case-Fatality Rates

Age-specific case-fatality rates (CFRs) were estimated by fitting a binomial model to the number of deaths (z_{ijk}) as a proportion of the number of total cases (y_{ijk}) in each age/sex/race bin. We denote the CFR for each group as ρ_{ijk}, so, $z_{ijk} \sim \text{Binomial}(y_{ijk}, \rho_{ijk})$.

Counterfactual Analysis of Mortality Disparities

To understand the relative importance of age-specific incidence versus case fatality as drivers of race/ethnic disparities as drivers of mortality disparities, we examined a pair of counterfactual scenarios in which: (1) age- and sex-specific COVID-19 incidence rates for each non-White race/ethnic group were replaced by the corresponding age-/sex-specific rate among White individuals, with original age- and sex-specific CFRs maintained and (2) the same procedure was repeated for CFRs, keeping race/ethnic incidence rates fixed for non-White individuals, while substituting White case-fatality for each non-White age/sex bin. We then used posterior simulation to obtain the difference in the number of deaths expected under each scenario to calculate the percentage reduction in observed deaths.

We wrote this carefully with goal of communicating the intuition behind this analysis, which is more sophisticated than relatively simple regression models we described above.

Software

All analyses were completed in R 4.0.3, using the rstanarm package [17] for Bayesian regression analysis, the tidybayes package for post-processing [18], and ggplot2 for visualization [19] (R Foundation for Statistical Computing).

RESULTS

In our dataset, there were 49 701 probable and confirmed COVID-19 cases and 5815 deaths attributable to COVID-19, for an overall CFR of 12%. Of these, 19 662 were among individuals identified as Black or African-American, 23 301 were among individuals identified as White, 1346 among individuals identified as Asian or Pacific Islander, 123 among individuals identified as Native American, and 1612 among individuals identified as belonging to any other racial/ethnic group in the 2018 ACS. Table 1 shows unadjusted per-capita case and mortality rates by race/ethnic group, as well as corresponding CFRs. Notably, the raw incidence rate among all non-White groups is substantially higher than among White individuals for all groups identified in the data except for Native Americans. However, the overall CFR for White individuals is on par with the CFR for Black individuals, potentially due to different distributions of ages among cases and deaths between these groups. Among White individuals, the average

If this paper was about counterfactual inference method, we would have included more detail on implementation. Since it is more about data and results, we kept discussion high-level.

We included this table to concisely communicate different facets of data (total cases, per-capita incidence rates, mortality rates, case-fatality rates, etc.) across different race/ethnic groups. A table is not the only good way to communicate this information, but we decided it would serve our purpose and was less time-consuming than constructing detailed figure.

Table 1. **Incidence, Mortality, and Demographic Characteristics of COVID-19 Cases and Deaths in Michigan, United States: March–June 2020 by Race/Ethnic Group**

Race	No. of Cases	No. of Cases/ 100 000	No. of Deaths	No. of Deaths/ 100 000	CFR, %	Avg. Age Years	Avg. Age at Death, Years	Female, %
Black	19 662	1445	2430	179	11	51	71	54
Latino	3657	734	133	27	2	38	67	48
Other	1612	626	104	40	5	45	72	51
Asjan / Pacific Islander	1346	440	76	25	5	44	77	52
White	23 301	311	3064	41	11	53	79	53
Native American	123	266	8	17	5	49	74	47

Abbreviations: Avg., average; CFR, case-fatality rate; COVID-19, coronavirus disease 2019.

This information is intended to enhance understanding of our results. It is not pivotal enough to our main conclusions to warrant its own figure, so it is included in text simply to provide reader a better sense about what results mean.

age of all reported cases was 53.4 years (95% posterior credible interval [CrI] = 53.2, 53.7), slightly older than among blacks (51.4 years; 95% CrI = 51.1, 51.6), and significantly older than among Latinos (38.1 years; 95% CrI = 37.6, 38.6) and those in the "other" race/ethnicity group. For all groups, the mean age among individuals with COVID-19 listed as their cause of death was significantly higher than for all cases within the same group. Among White individuals, the average age at death was greatest, at 79.2 years (95% CrI = 78.6, 79.9), 8 years higher than among Black individuals at 71.2 years (95% CrI = 70.5, 71.9), with Latinos having the youngest average age at death at 66.7 years (95% CrI = 63.6, 69.8).

Standardized Incidence and Mortality Rates

Table 2 contains age- and sex-standardized incidence and mortality rates per 100 000 population and corresponding between-group rate ratios, by race/ethnic group. Rows of the table are ordered by raw incidence per 100 000 individuals for comparability with Table 1. This shows that the general patterns in the raw incidence and mortality hold after adjustment, although the age- and sex-adjusted incidence among Latinos increased, reflecting the younger average age of cases identified as Latino. The provided incidence rate ratios (IRRs) and mortality rate ratios (MRRs) show the enormous disparity in incidence and mortality between Black and White individuals, with an IRR of 5.5 and an MRR of nearly 7. Again, these IRRs and MRRs reflect the fact that all groups other than Native Americans had higher rates of incidence and mortality than White individuals and that these differences do not simply reflect the age and sex distribution of cases. For Native Americans, rates were statistically indistinguishable from those for White individuals, although this may be due to the very small number of cases and deaths overall in this group in our data. In the following sections, we will examine age-stratified incidence and mortality rates by race/ethnicity for Black, Latino, Asian/Pacific Islander and White individuals. Native Americans are excluded from age-stratified analyses due to a small sample size, as are individuals in the "Other" race/ethnic categorization.

Cumulative Incidence Rates

Figure 1 illustrates the dramatically higher overall and age-specific incidence rates among Black individuals and individuals in the "Other" race/ethnic category than for White individuals, particularly at older ages at which individuals are far more likely to die from their infection. In addition, the horizontal dashed line in each panel of Figure 1 shows the raw incidence rate for each group. The extent of these disparities in incidence is clearly in evidence in the left-hand panel of Figure 3, which shows the ratio of the age-specific cumulative incidence rate (IRR) for each race/ethnic group as compared with the comparable rate for White individuals. In this case, rates for all non-White groups are significantly higher, with these disparities most pronounced at older ages for Black individuals and younger ages for Latinos. The IRR for individuals in the "other" group was fairly consistent across ages, with a small drop in the 20–40-year age range.

Table 2. **Age- and Sex-Standardized COVID-19 Incidence and Mortality Rates and Corresponding Rate Ratios, by Race/Ethnic Group, in Michigan, United States: March–June 2020**

Race	Incidence/100 000	IRR	Mortality/100 000	MRR
Black	1626 (1602, 1649)	5.5 (5.4, 5.6)	244 (234, 255)	6.7 (6.4, 7.1)
Latino	912 (879, 946)	3.1(3, 3.2)	69 (57, 82)	1.9 (1.6, 2.3)
Other	1150 (1088, 1216)	3.9 (3.7, 4.1)	123 (99, 149)	3.4 (2.7, 4.1)
Asian/Pacific Islander	498 (469, 529)	1.7 (1.6, 1.8)	51 (41, 64)	1.4 (1.1, 1.8)
White	297 (293, 300)	Ref	36 (35, 38)	Ref
Native American	285 (237, 341)	1 (0.8, 1.2)	29 (13, 54)	0.8 (0.3, 1.5)

The table shows incidence rates and mortality rates and 95% CrI, as well as corresponding standardized IRRs and MRRs. For all ratio measures of association, the incidence and mortality rate among White individuals is used as the reference group. Abbreviations: COVID-19, coronavirus disease 2019; CrI, posterior credible interval; IRR, incidence rate ratio; MRR, mortality rate ratio; Ref, reference.

We included these population-level results as a table because: 1) rows and columns map onto those presented in Table 1 and highlights disparities are confirmed by statistical modeling results; 2) simple and 'flat' presentation of model results here clearly differentiates it from detailed, age-specific results presented in Figures 1-3. Structuring things this way provides more detail than can be communicated in simple results in Table 1.

Much like with the tables, we ensured that presentation of information in Figures 1-3 “rhymed” in a way that made them easy to compare and reason about. Each figure is organized in terms of age-specific incidence, case-fatality, or disparities in each. This is meant to make it easy to go back and forth between the figures and understand what they mean with respect to each other.

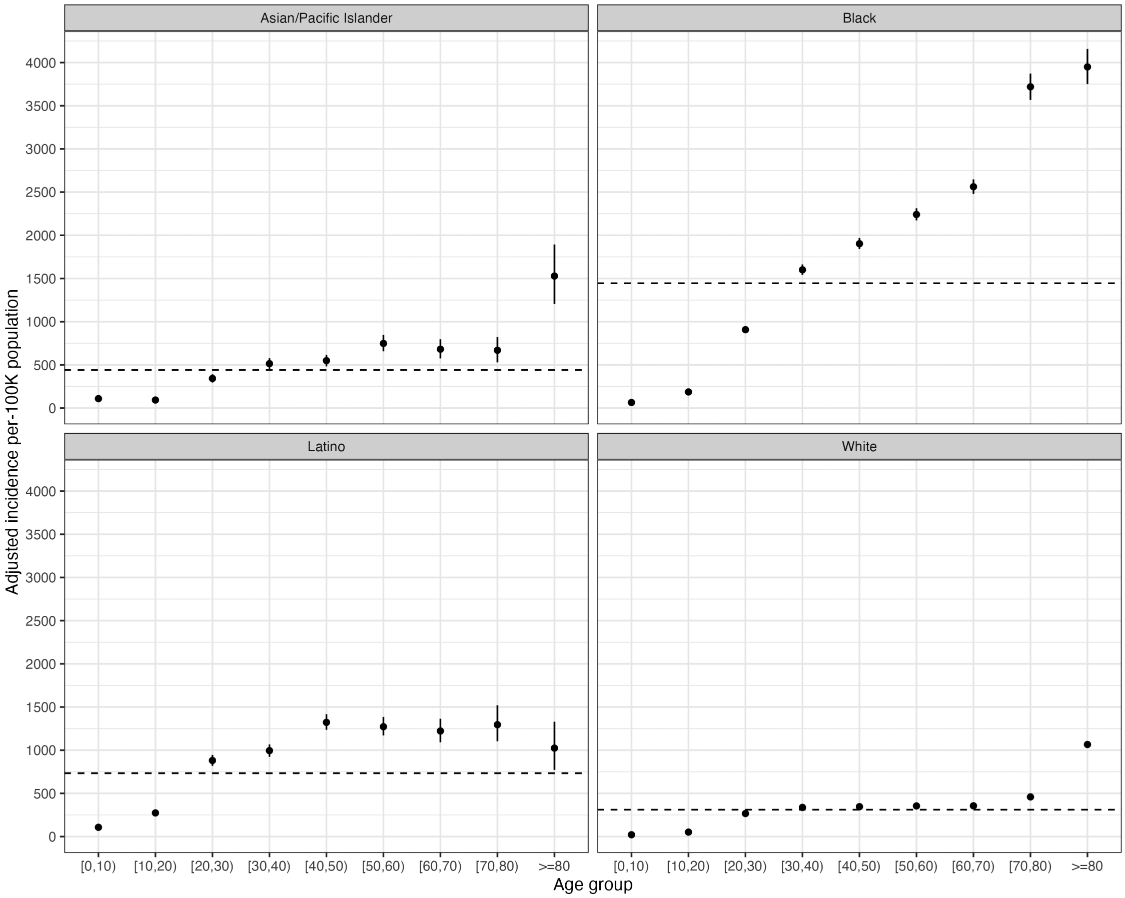

Figure 1. Incidence rate estimates (points) and 95% CrI (vertical lines) of COVID-19 infection per 100 000 population by 10-year age groups, stratified by race/ethnic group. Dashed lines indicate the crude rate for each group. Abbreviations: COVID-19, coronavirus disease 2019; CrI, posterior credible interval.

Case-Fatality Rates

Figure 2 illustrates a steadily increasing trend in the probability of death among identified cases from age 50 onwards across groups, although there are differences in these rates at younger ages. These are visible in the right-hand panel of Figure 3, which shows the ratio of the age-specific CFR for Black individuals, Latinos, and those in the “other” group versus White CFRs. Because of the small number of deaths among individuals younger than 20 years of age, these groups are excluded from the figure. For Black individuals, all age groups from 30 to 70 years experienced higher CFRs than White individuals, with this disparity most pronounced among 40- to 49-year-olds. However, for Latinos and those in the “other” race/ethnic group, there are no significant differences in age-specific CFRs as compared with White individuals. These results and those in Table 2 suggest that, although there are meaningful differences in case fatality by race and age, the large disparities in COVID-19 mortality cannot be explained by CFRs alone.

Counterfactual Analysis of Mortality Disparities

We found that substituting the incidence rates of White individuals for those of non-White individuals would result in a decrease of 82% (95% CrI = 81%, 84%) of the observed deaths among Blacks individuals, 57% (95% CrI = 47%,66%) among Latinos, and 35% (95% CrI = 18%,49%) among Asian/Pacific Islanders. In the second scenario (ie, when White CFRs were substituted for non-White CFRs but group-specific incidence rates maintained), we found no significant change in the expected number of deaths for any group except for Blacks individuals, who saw a smaller but still meaningful decrease of 19% (95% CrI = 14%,25%) of deaths.

These results suggest that, while differential CFRs can account for some of the disparity in Black versus White mortality rates, the large majority of COVID-19 deaths among African-Americans in Michigan can be attributed to the large differences in age-specific incidence illustrated in Figure 1. Similarly, although Latinos and Asian/Pacific Islanders had similar crude

Sensitivity analysis lets us understand the limitations of our results in greater depth and can also increase our confidence in them. From a pragmatic perspective, it also signals to reviewers and editors that you have thought about objections an informed reader might have and have tried to understand and address them, which is important for retaining their trust in your analysis and interpretation.

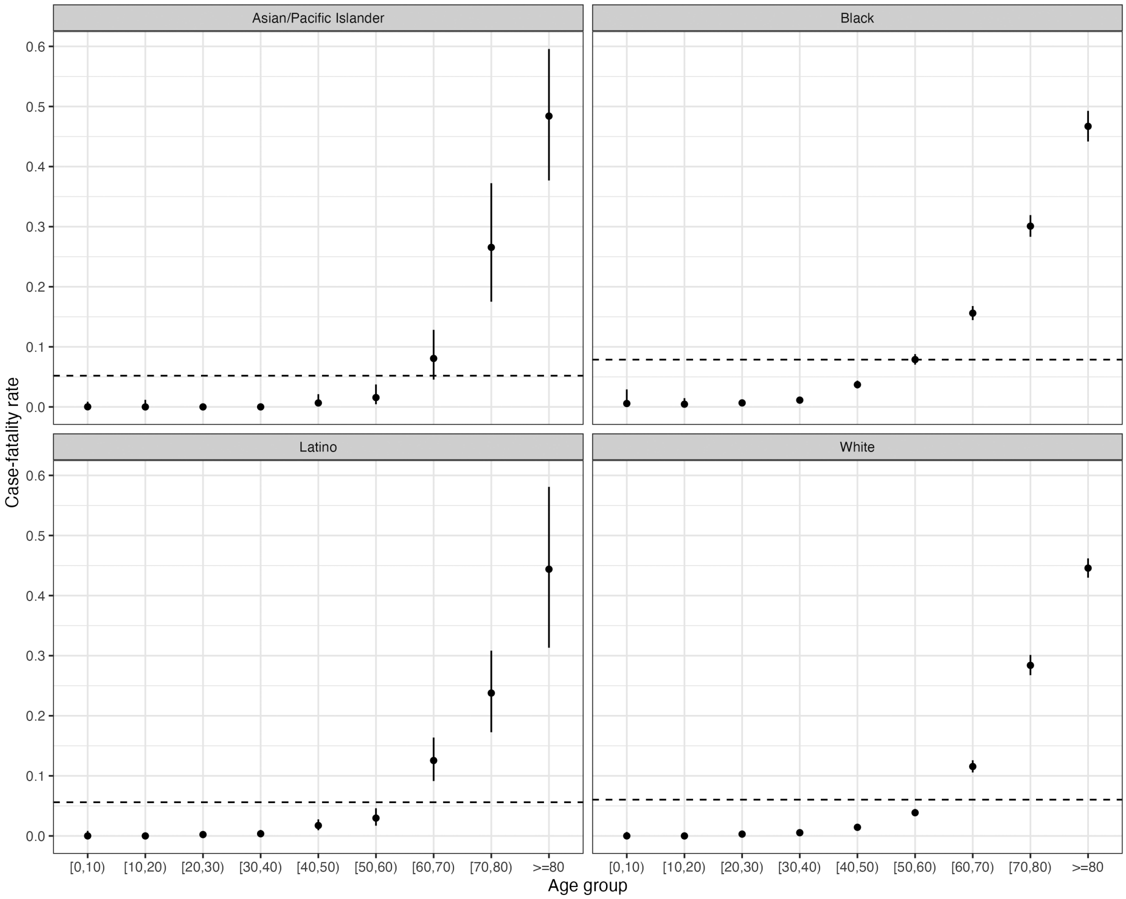

Figure 2. COVID-19 case-fatality rate estimates (points) and 95% CrI (vertical lines) by 10-year age groups, stratified by race/ethnic group. Dashed lines indicate the crude rate for each group. Abbreviations: COVID-19, coronavirus disease 2019; CrI, posterior credible interval.

mortality rates to White individuals (Table 1), these results indicate that these rates would be significantly lower if their exposure risks were more similar to their White peers.

Sensitivity to Case Definition

To ensure that our results were not strongly impacted by the combined analysis of probable [15] and PCR-confirmed SARS-CoV-2 cases, we conducted a sensitivity analysis in which all results were re-generated using data from only PCR-confirmed cases. Descriptive analysis showed that younger and White individuals were more likely to have a probable infection than older and non-White individuals. When these probable cases were excluded, incidence and mortality disparities for younger non-White individuals increased, but our results for older individuals—who experienced the bulk of mortality—remain qualitatively unchanged, as do our population-level conclusions. For full results of this analysis, see the Supplementary Materials.

DISCUSSION

Our results highlight large gaps in COVID-19 incidence and mortality in Michigan that cannot be explained by differences in population age and sex composition. Results from our counterfactual analysis suggest that the stark differences in crude and adjusted mortality between Black individuals and all other race/ethnic groups shown in Tables 1 and 2 are driven in large part, but not exclusively, by disparities in infection risk at all ages, particularly an extremely high rate of COVID-19 infection among older Black individuals in particular. This group had a CFR similar to same-aged White individuals, but reported infection rates 6–8 times greater than their White counterparts. Some of this disparity is also driven by the higher CFR among middle-aged Black people, as compared with same-aged White individuals, in combination with the 5–6 times greater risk of infection among middle-aged Black individuals as compared with White individuals.

This is as concise a summary of the major results and their implications as we could construct. We hope that if a reader only read this paragraph, they would get why the work is important and what its broad implications are.

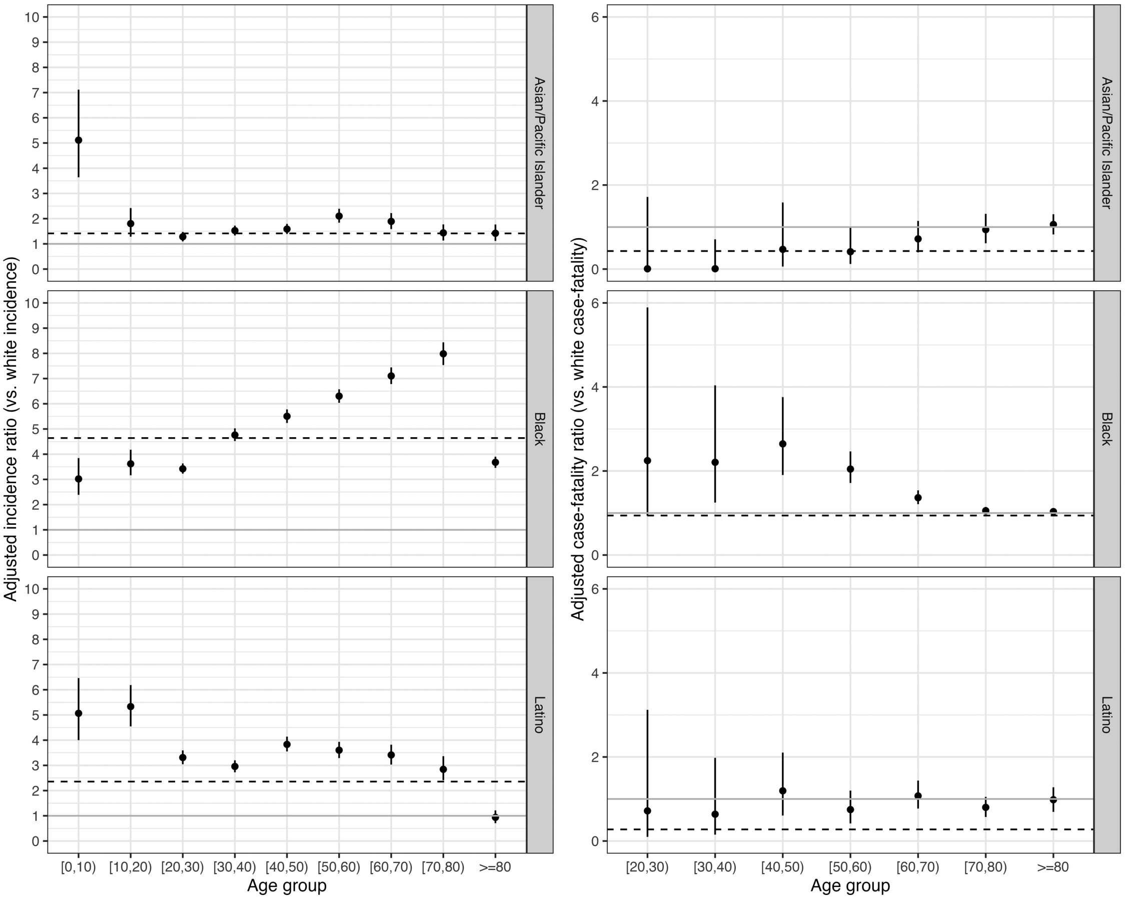

Figure 3. Disparities (as measured by rate ratios [points]) and 95% CrI (vertical lines) of COVID-19 incidence (left-hand column) and case-fatality rate (right-hand column) by age and race/ethnic group compared with White individuals. Dashed lines indicate the ratio of the crude overall rate for each group; the solid gray line is a guide for assessing the strength of association, representing a rate ratio of 1.0 (no association). Abbreviations: COVID-19, coronavirus disease 2019; CrI, posterior credible interval.

Despite these unambiguous results, the full extent of racial and socioeconomic disparities in COVID-19 outcomes in Michigan and the rest of the United States is likely to be even worse than is reflected in administrative data of the type analyzed here. Results from other hard-hit cities, states, and countries have indicated high rates of excess mortality reflective of unrecognized and unreported COVID-19 infection [20]. In a recent analysis of excess mortality using state-level data, Weinberger et al [21] found that there were approximately 4700 unreported deaths likely due to COVID-19 or another respiratory infection in the United States during the period from 1 March to 30 May 2020, for a rate of 61 per 100 000 unreported deaths from COVID-19 above reported totals. In addition, the damage to health from the pandemic goes beyond the direct impact of infections and deaths from SARS-CoV-2. For example, Woolf et al [22] showed that 33% of the total excess deaths during the period from 1 March to 25 April 2020 in Michigan were attributable to noninfectious causes, with the remainder associated with respiratory infections, primarily COVID-19. Although these results are not broken down by race/ethnicity, it is likely that the burden of mortality is not equally shared across race/ethnic groups and socioeconomic strata. Beyond delays in healthcare seeking due to the pandemic, it is quite likely that these patterns of excess death reflect underlying disparities in chronic illnesses that predispose individuals to mortality from COVID-19, lack of access to healthcare for Black and Latinx individuals and other minority groups, and variable quality of care delivered based on racial/ethnic identity.

Here, we connect our analysis to broader problem of racial disparities in infectious disease outcomes to show why our work is significant and may have utility that extends beyond topic at hand.

When interpreting these and other results illustrating racial disparities in COVID-19 incidence and mortality, it is key not to portray race as a risk factor independent of health conditions, wealth, and other potentially modifiable risk factors [23] that may predispose individuals to COVID-19 infection and death.

This is a kind of hidden limitations statement: We used race/ethnicity as predictors of disease outcomes but want readers to be clear that we understand that these are not causal variables on their own but instead reflect a complex mix of social and economic factors.

An important limitation: Our analysis is about data we had access to and may not generalize. But this paragraph is also a classic of the “yes, but” genre of limitations statements because we then get into why this kind of generalization may be reasonable even if we can’t be sure.

Decision to make opening sentences of conclusion blunt and focused on necessity of action to address problem was deliberate and reflects that this paper was written and published at height of COVID-19 pandemic. Had it been written later, implications of our results and call to action in conclusion would be different and perhaps more measured.

For example, McClure et al [24] illustrate how a focus on—and adjustment for—individual-level “underlying conditions” obscures the role of racial inequality in shaping the prevalence of these chronic health conditions and other factors such as residence in multigenerational households, which may increase risk among racial and ethnic minority groups.

A strength of our analysis is the use of detailed case data obtained directly from the MDSS. This allowed us to identify age- and race-specific risks of COVID-19 infection and death. Nonetheless, there are some limitations that are important to highlight. First, our reliance on census-defined race/ethnicity as a proxy for exposure and mortality risk is necessarily reductive and does not shed light on factors that can be modified to reduce these disparities [25]. Future analyses are necessary using either prospectively collected data inclusive of SES or spatial analyses that join neighborhood-level information on wealth and other markers of SES with individual-level case data. The set of cases obtained from MDSS during this period is also necessarily incomplete, with large numbers of asymptomatic and less-severe infections undoubtedly missing from this registry.

In addition, although the disparities in our data likely mirror those nationwide, it is important to remember that these results reflect patterns of infection and death in Michigan during the first wave of the COVID-19 pandemic. Although its relatively large population size and socioeconomic and racial composition make Michigan a bellwether of many national trends, this analysis should be interpreted relative to its context. Consequently, similar analyses are sorely needed to understand how these outcomes vary across locales and populations. For a complete discussion of these issues, see the Supplementary Materials.

Because of the deep structural roots of the disparities identified in this analysis, it is easy—but wrongheaded—to conclude that there is nothing to be done. The fluid nature of the COVID-19 pandemic and its response provides opportunities to narrow these appalling inequities in infection and death, particularly as new therapeutics and vaccines against SARS-CoV-2 become available. For this to be the case, however, similar amounts of effort to what has been done to open college campuses and other workplaces need to be focused on increasing the quality and quantity of testing, healthcare, and social support among people of color. While understanding the causes of disparate outcomes is important, it does not necessarily instruct us on what to do. If the current pandemic teaches us something, it is that closing the gap in infection and mortality during the current catastrophe—and preventing such inequities in the next one—requires addressing the racialized dismantling of public infrastructure and systematic divestment that has made these disparities in exposure, susceptibility, and mortality a foregone conclusion [26]. Accomplishing this necessitates an urgent re-orientation around an “epidemiology of consequence” [27] that can identify and attack the structural and practical barriers to health equity before the next disaster strikes.

Supplementary Data

Supplementary materials are available at *Clinical Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Notes

Author contributions. J. Z. and R. T. conceptualized and completed data analysis; J. Z., R. T., R. N., A. C., N. M., and P. D. participated in data preparation and cleaning; all authors participated in the writing and editing of the final manuscript.

Acknowledgments. The authors gratefully acknowledge the efforts and insight of Sarah Lyon-Callo, Jim Collins, and other members of the Michigan Department of Health and Human Services, as well as the input of Emily Martin, Marisa Eisenberg, and Sharon Kardia at the University of Michigan.

Disclaimer. The funding sources had no role in the preparation of this manuscript.

Financial support. J. Z., R. T., and A. C. were supported by a COVID-Pursuing Original Data Science (PODS) grant from the Michigan Institute for Data Science at the University of Michigan. R. T. was supported by the MCubed program at the University of Michigan. J. Z., R. M., and N. M. were supported by award 6 U01 IP00113801-01 from the US Centers for Disease Control and Prevention.

Potential conflicts of interest. The authors: No reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. DiMaggio C, Klein M, Berry C, Frangos S. Blacks/African Americans are 5 times more likely to develop COVID-19: spatial modeling of New York City ZIP code-level testing results. Epidemiology **2020**; 51:7–13.
2. Mahajan UV, Larkins-Pettigrew M. Racial demographics and COVID-19 confirmed cases and deaths: a correlational analysis of 2886 US counties. J Public Health **2020**; 42:445–7.
3. Patel AP, Paranjpe MD, Kathiresan NP, Rivas MA, Khera AV. Race, socioeconomic deprivation, and hospitalization for COVID-19 in English participants of a national biobank. Int J Equity Health **2020**; 19:114.
4. Rodriguez-Lonebear D, Barceló NE, Akee R, Carroll SR. American Indian reservations and COVID-19: correlates of early infection rates in the pandemic. J Public Health Manag Pract **2020**; 26:371–7.
5. McLaren J. Racial disparity in COVID-19 deaths: seeking economic roots with census data. Natl Bureau Econ Res **2020**; doi:10.3386/w27407.
6. Moran E, Kubale J, Noppert G, Malosh RE, Zelner JL. Inequality in acute respiratory infection outcomes in the United States: a review of the literature and its implications for public health policy and practice. medRxiv **2020**. doi: 10.1101/2020.04.22.20069781.
7. Saloner B, Parish K, Ward JA, DiLaura G, Dolovich S. COVID-19 cases and deaths in federal and state prisons. JAMA **2020**; 324:602–3.
8. Kajeepeta S, Rutherford CG, Keyes KM, El-Sayed AM, Prins SJ. County jail incarceration rates and county mortality rates in the United States, 1987. Am J Public Health **2020**; 110:S109–15.
9. Acevedo-Garcia D. Residential segregation and the epidemiology of infectious diseases. Soc Sci Med **2000**; 51:1143–61.
10. Roberts SK. Infectious fear: politics, disease, and the health effects of segregation. Chapel Hill, NC: University of North Carolina Press, **2009**.
11. Yancy CW. COVID-19 and African Americans. JAMA **2020**; 323:1891–2.
12. Meyerowitz-Katz G, Merone L. A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates. medRxiv **2020**. doi:10.1101/2020.05.03.20089854.
13. Nepomuceno MR, Acosta E, Alburez-Gutierrez D, Aburto JM, Gagnon A, Turra CM. Besides population age structure, health and other demographic factors can contribute to understanding the COVID-19 burden. Proc Natl Acad Sci U S A **2020**; 117:13881–3.
14. Millett GA, Jones AT, Benkeser D, et al. Assessing differential impacts of COVID-19 on black communities. Ann Epidemiol **2020**; 47:37–44.
15. Michigan Department of Health and Human Services. Michigan state and local public health COVID-19 standard operating procedures. Lansing, MI: Michigan Department of Health and Human Services, **2020**:41.
16. Ruggles S, Flood S, Goeken R, et al. IPUMS USA: version 10.0. **2020**.

Supplement includes nuanced ‘side’ analyses helpful for interpreting major conclusions, but not essential for understanding what we did, why we did it, and what it means.

We ended paper by emphasizing that big structural inequalities highlighted by our analysis require big, structural changes if they are to be addressed in a meaningful way. Even though this is not a paper about interventions per se, we wanted to be sure that our perspective on what should be done, and timeline over which we can expect to see meaningful change, is clearly reflected in text.

17. Gabry J, Ali I, Brilleman S, et al. Rstanarm: Bayesian applied regression modeling via Stan. **2020**. Available at: https://CRAN.R-project.org/package=rstanarm. Accessed 14 July 2020.
18. Kay M, Mastny T. Tidybayes: tidy data and "Geoms" for Bayesian models. **2020**. Available at: https://CRAN.R-project.org/package=tidybayes. Accessed 14 July 2020.
19. Wickham H, Chang W, Henry L, et al. Ggplot2: create elegant data visualisations using the grammar of graphics. **2020**. Available at: https://CRAN.R-project.org/package=ggplot2. Accessed 14 July 2020.
20. Olson DR, Huynh M, Fine AD, et al.; New York City Department of Health and Mental Hygiene (DOHMH) COVID-19 Response Team. Preliminary estimate of excess mortality during the COVID-19 outbreak New York City, March 11–May 2, 2020. MMWR Morb Mortal Wkly Rep **2020**; 69:603–65.
21. Weinberger DM, Chen J, Cohen T, et al. Estimation of excess deaths associated with the COVID-19 pandemic in the United States, March to May 2020. JAMA Intern Med **2020**; 180:1336–44.
22. Woolf SH, Chapman DA, Sabo RT, Weinberger DM, Hill L, Taylor DDH. Excess deaths from COVID-19 and other causes, March-July 2020. JAMA **2020**; 324:1562–4.
23. Sen M, Wasow O. Race as a bundle of sticks: designs that estimate effects of seemingly immutable characteristics. Ann Rev Polit Sci **2016**; 19:499–522.
24. McClure ES, Vasudevan P, Bailey Z, Patel S, Robinson WR. Racial capitalism within public health: how occupational settings drive COVID-19 disparities. Am J Epidemiol **2020**; 189:1244–53.
25. Williams DR, Collins C. Racial residential segregation: a fundamental cause of racial disparities in health. Public Health Rep **2001**; 116:404–16.
26. Bailey ZD, Moon JR. Racism and the political economy of COVID-19: will we continue to resurrect the past? J Health Polit Policy Law **2020**; 45:937–50.
27. Keyes K, Galea S. What matters most: quantifying an epidemiology of consequence. Ann Epidemiol **2015**; 25:305–11.