Lung (2022) 200:531–537
https://doi.org/10.1007/s00408-022-00559-6

EDITORIAL

The Statistics of Counting Coughs: Easy as 1, 2, 3?

Matthew Rudd[1] · Woo-Jung Song[2] · Peter M. Small[3]

Received: 27 July 2022 / Accepted: 28 July 2022 / Published online: 16 August 2022

Cough is one of the most common reasons that individuals seek health care, and yet it is largely unmeasured in clinical medicine and practice. New technology that unobtrusively monitors cough holds great promise in continually monitoring cough. Detecting a change in cough rates would be easy if people coughed like metronomes. In reality, chronic coughers have good and bad days, hours, and minutes. This stochastic nature of cough raises challenges in detecting statistically significant changes in cough frequency. Here we describe statistical properties of cough monitoring data and suggest a method to detect changes in its frequency.

Cars, Coughs, and Capricious Chance

Suppose you want to understand traffic patterns where you live, maybe to help local city planners or simply to satisfy your curiosity. A natural approach would be to find a spot on the side of the road and just start counting cars. But where should you go and when? Traffic varies a lot by time of day and location—observing a busy thoroughfare during rush hour will give a much different impression than watching a quiet neighborhood on a lazy afternoon. Even watching the same location at the same time for several days in a row could yield very different numbers, for no other reason beyond the caprices of chance. But watching that same location at that same time for months or years would reveal the structure behind those seemingly random counts. You would know, for example, just how likely it would be to see 14 or 40 or 4000 cars pass by that location over the course of an hour.

This all makes sense based on our common experiences, as people are familiar with the ebb and flow of daily life. Some of the rhythms around us are more elusive, though, so routine as to be hidden in plain sight—or sound—in the case of cough. Cough is one of the most common reasons people seek medical care [1, 2], and yet to date it is simply not measured continuously [3] leaving unanswered important questions such as: How many times did you cough yesterday? When did you cough the most? How long did you go without coughing at all? If you're not sure, you're hardly alone. And if your cough patterns change, how do you know? How can you find what triggers your cough? What do you tell your doctor? Is a new medication improving your cough?

Monitoring Matters

Without understanding the randomness of cough, it would be impossible to resolve questions about hourly cough and patterns objectively and scientifically. With a bit of statistical insight and code, however, some answers are readily available and desperately needed. Even though cough is one of the most common complaints that leads people to seek medical care, patients and doctors quickly reach an impasse when they meet. People generally lack both the vocabulary to describe coughs and the ability to quantify their coughs over time. While the common diagnostic tools that we take for granted, such as thermometers, blood pressure monitors, and routine blood tests, have enabled a data-driven revolution in medical care, the fundamental elements of cough frequency measurement remain mostly unchanged since the 1960s [4]. What are "normal" cough patterns for healthy people, adults with tuberculosis [5], or children with malaria? What changes in cough signify the start of a particular illness or

✉ Matthew Rudd
matthew.r@hyfe.ai

Woo-Jung Song
swj0126@gmail.com

Peter M. Small
peter@hyfe.ai

[1] Department of Mathematics and Computer Science, The University of the South, Sewanee, TN, USA

[2] Department of Allergy and Clinical Immunology, Asan Medical Center, University of Ulsan College of Medicine, 88, Olympic-ro 43-gil, Songpa-gu, Seoul, Korea

[3] Research and Development Department, Hyfe Inc, Wilmington, DE, USA

an exacerbation of a condition like asthma, COPD, or lung cancer? Nobody knows for the most mundane of reasons: not enough good data.

The first step towards a better understanding of cough is diligent data collection: track all coughs as they occur and maintain records that can be carefully analyzed [6, 7]. What we are learning is that coughs are like traffic on city streets. Like busy streets, some people cough a lot, even more at certain times than at others; and like quiet cul-de-sacs, some people do not cough very much at all regardless of the time of day. And just as the number of cars passing by a given spot at a certain time can vary considerably from day-to-day, the number of times a person coughs during a given hour can change quite a bit from one day to the next, either purely by chance or in response to some internal or external change or treatments [8]. Distinguishing between these possibilities is a basic goal of cough science and one that can only be achieved after fully understanding the inherent randomness of coughing.

Data!

To make this concrete, let's look at some real data—one person's coughs, tracked continuously with a smartphone-based automated AI cough monitor—Hyfe Cough Tracker (version ar1.7.1(1)), over the same 4 h period, from 11 a.m. to 3 p.m., on 3 consecutive days (Fig. 1). The data were collected with approval by the Institutional Review Board of the Asan Medical Center (IRB No. 2021-1632). Informed consent was obtained from the individual to be included in this manuscript. The procedures used in this study adhere to the tenets of the Declaration of Helsinki. Some patterns out of this data are as follows:

- This person coughed a lot—35 times—between 11 a.m. and noon on February 20, but coughed far less during that hour on the 21st and even less on 22nd.
- This person did not cough much—only five times—between 2 p.m. and 3 p.m. on February 20, but coughed more during that hour on the 21st and even more on 22nd.
- Between noon and 2 p.m., this person's cough counts were low on 20, much higher on the 21st, then low again on 22nd.

In other words, there is no consistent pattern here! These are simply the vagaries of chance in action: individual outcomes in any given hour can easily change dramatically and cannot be predicted exactly ahead of time. This does *not* mean this person's coughing lacks structure but rather we just need more data to unlock that structure and make sense of it.

More Data!

Instead of just looking at 12 observations (four one-hour periods over three days), let us look at 895 hourly observations of this person's coughs collected over the course of two months through a passive and unobtrusive smart device-based listening process (Fig. 2).

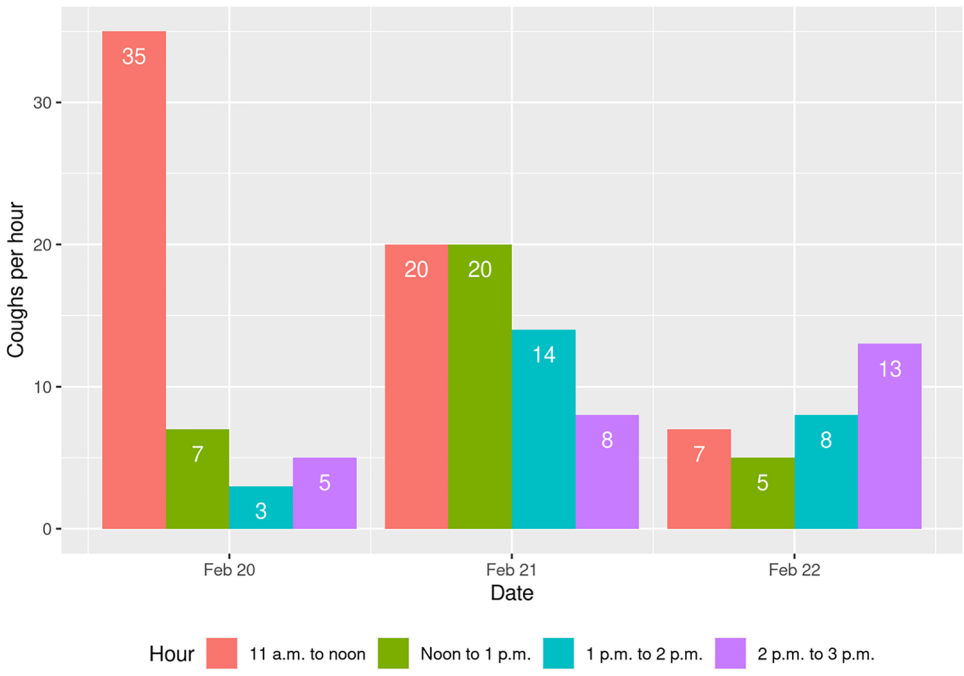

Fig. 1 4 h long continuous cough data of a chronic cougher over 3 consecutive days

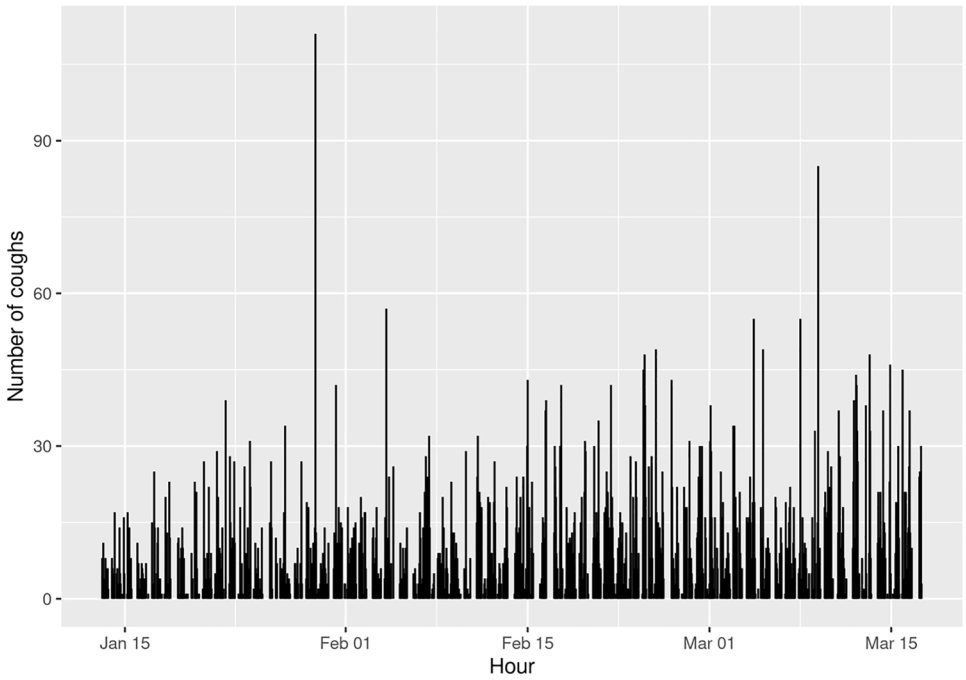

Fig. 2 Continuous cough monitoring data of the same chronic cougher as in Fig. 1, collected over the course of 2 months

From this perspective, the 3 days considered earlier in Fig. 1 (February 20-22) are unremarkable. In fact, this person's coughs look fairly stable throughout this 2-month period, with some clear and regular undulations. A couple of individual hours do jump off the page; however this person coughed:

- 111 times between 4 p.m. and 5 p.m. on January 29, and then
- 85 times between 9 a.m. and 10 a.m. on March 9.

What should we make of these two counts? Are they meaningful aberrations, perhaps indicating the onset of an illness, or are they just due to random chance, signifying nothing in particular? To find out, we need to look at the data in a different way. Instead of organizing the hours chronologically as in Fig. 2, let us group them by the number of coughs per hour to see how often the different counts occur (Table 1). This can be plotted in a table that goes on for pages, but let us just look at the first bit of data.

The most likely outcome is just one cough per hour. This happened 88 times, or 9.83% of the 895 h available. (There are 620 h when either 0 coughs were recorded or monitoring was not being done. Since we cannot determine which is the case for those missing hours, we are omitting them from this analysis.) About 50% of the time this person coughed fewer than 8 times per hour. Paging forward through this table shows hourly counts exceeding 30 were rare. This person coughed 32 times twice, 35 times once, and 40 times never. Rather than looking up all possible individual counts, however, we should compute percentages for *ranges* of values, for example this person coughed 30 times or more 5.7% of the time. This proportion naturally decreases as we increase the number of coughs; there were 48 or more coughs in just 1% of the hours.

Table 1 Grouping continuously monitored coughs by the number of coughs per hour

Number of coughs (*N*)	Number of hours with *N* coughs	Percentage of hours with *N* coughs
1	88	9.83%
2	76	8.49%
3	74	8.27%
4	55	6.15%
5	46	5.14%
6	60	6.70%
7	48	5.36%
8	48	5.36%
9	40	4.47%
10	33	3.69%

Modeling Hourly Counts

A histogram in Fig. 3 summarizes all of this nicely, saving us the hassle of scrolling through a large table (but at the cost of making it harder to compute precise percentages). Each of the bars in Fig. 3 represents one possible number of coughs, as indicated on the horizontal axis; the height of each bar is the number of hours during which

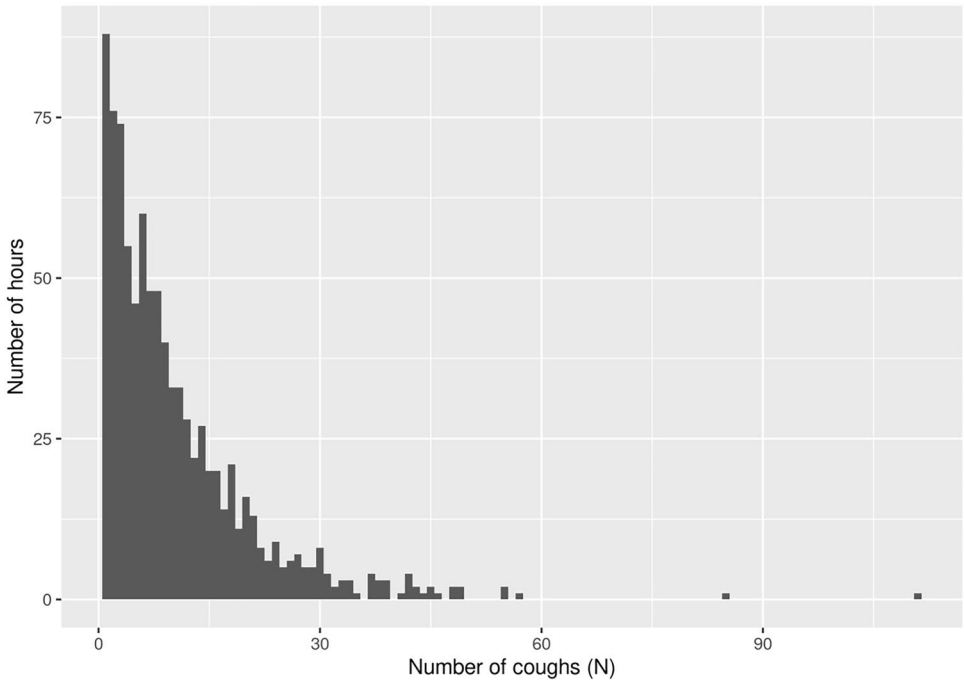

Fig. 3 Continuously monitored coughs by the number of coughs per hour. Each of the bars on the horizontal axis represents one possible number of coughs; the height of each bar is the number of hours during which that particular number of coughs occurred

that particular number of coughs occurred. The two unusual outcomes, 111 and 85, are easy to spot. Thanks to an abundance of data, this histogram has some obvious structure, strongly skewed to the right, with small counts much likelier than large counts. But we would need far more data to have a nice smooth histogram without such a jagged profile.

What would this histogram look like if we had several *years* of hourly observations? Instead of waiting to collect all that data, just ask a statistician. One look at this histogram is all it takes to recognize an old friend, a famous distribution of counts known as the *negative binomial distribution*. It also looks like the even more famous Poisson distribution—some additional calculations are needed to know which one we have got here. A Poisson distribution's average and variance are exactly equal, while a negative binomial distribution's variance exceeds its average, making it overdispersed. It turns out that cough counts are almost always overdispersed, as these are.

This distribution has a precise mathematical formula [9], but is essentially a giant table that fills in all of the blanks above, providing the theoretical probability of each possible outcome, not just the ones we happen to have observed thus far. Let us add these *theoretical percentages* and the corresponding *theoretically expected counts* to our table of observed results to see how they are compared (Table 2).

As mentioned earlier, we have an issue with missing data. We do not know how to distinguish between hours with 0 coughs and hours without monitoring, but we would expect the number of zeros to be close to the number of ones or

Table 2 Grouping continuously monitored coughs by the number of coughs per hour with the addition of theoretically expected cough counts

Number of coughs (*N*)	Number of hours with *N* coughs	Percentage of hours with *N* coughs	Theoretical number of hours with *N* coughs	Theoretical percentage of hours with *N* coughs
0	0	0%	64.54	7.21%
1	88	9.83%	65.51	7.32%
2	76	8.49%	62.86	7.02%
3	74	8.27%	59.14	6.61%
4	55	6.15%	55.1	6.16%
5	46	5.14%	51.02	5.70%
6	60	6.70%	47.06	5.26%
7	48	5.36%	43.28	4.84%
8	48	5.36%	39.72	4.44%
9	40	4.47%	36.39	4.07%
10	33	3.69%	33.29	3.72%

twos. While this undoubtedly affects the fit between the observed and expected percentages, they still agree quite well overall. This is even easier to see if we plot the theoretically expected counts along with the histogram of observations (Fig. 4).

How do we nail down the specific probability distribution shown in Fig. 4 in blue, namely, the negative binomial distribution that fits this data? It is determined by two statistics, the average and the standard deviation of this person's coughs per hour, which happen to be 10.55 and 10.47, respectively in our described case. We have good estimates of these statistics for this particular person, since we have 895 hourly observations, but getting good estimates in other situations presents significant modeling challenges.

Simulations and Aberrations

Remember the questions posed earlier regarding whether the two outstandingly large hourly counts, 111 and 85, are noteworthy? Could they just happen by chance? Now that we have a good theoretical model of this person's general cough pattern, we can answer these questions by simulating what might be observed at other times. For example, if we monitored this person for six different sessions, each lasting 895 h, statistics show that we could end up with the following histograms of hourly counts (Fig. 5)—whose maxima vary quite a bit (Table 3).

One of these six simulated datasets in Fig. 5 has a maximum larger than this person's maximum of 111 coughs in one hour (Fig. 3), but this does *not* mean that there is a one out of six chance of observing such a large hourly count. The chance is actually much smaller; in 10,000 simulated monitoring sessions of length 895 h, the maximum hourly count was 111 or larger 207 times, 2.07% of the simulations. This means that, for our described person, a maximum of 111 coughs (Fig. 3) in one out of 895 h is indeed unusual, but far from impossible (Fig. 5). In those same 10,000 simulations, 21.18% of the maximum hourly counts were 85 or larger, making an observation of that size quite likely. In short, these two outliers are probably just due to chance, *not* indications of significant changes to this person's baseline cough pattern.

Conclusions

Knowing if a cough count or pattern represents a noteworthy change of state or is just a statistical anomaly has, up until now, not been a question cough science could answer. Continuous, unobtrusive cough monitoring over sufficiently long periods of time is the solution to this problem of detecting change in cough rates. Some people will see immediate results. Chronic coughers who have felt dismissed by their doctors will finally be able to show them just how much they cough and get the care they deserve. Other people will reap the benefits of monitoring once science advances a bit further. People whose cough patterns change significantly will find out in real time and be able to seek appropriate care faster than ever before. The available care will also improve, as researchers working on new cough treatments

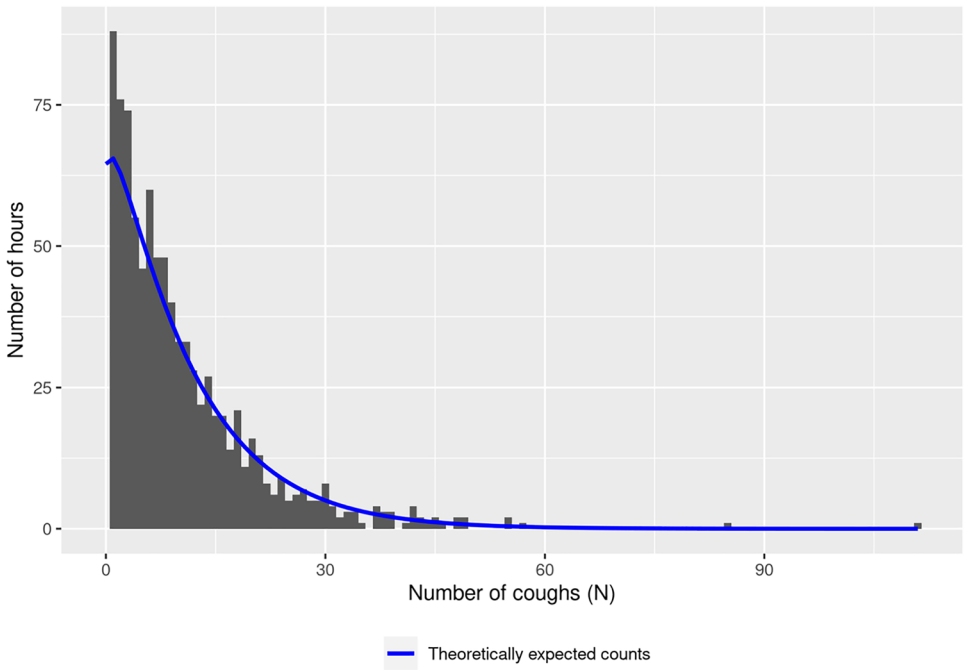

Fig. 4 Continuously monitored coughs by the number of coughs per hour with the addition of theoretically expected cough counts

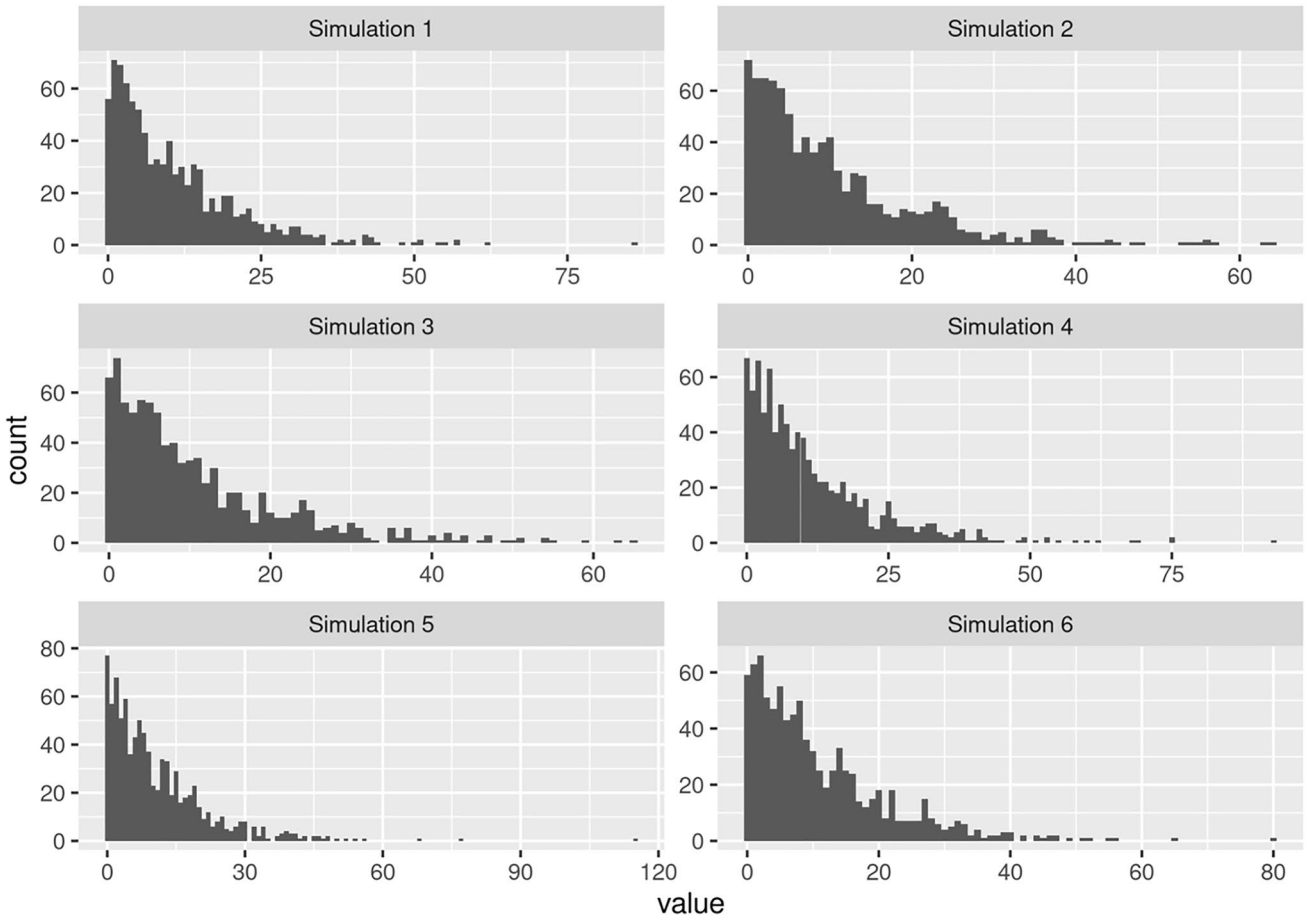

Fig. 5 Simulated scenarios in six different sessions (each 895 h of continuous cough monitoring) and hourly cough counts

Table 3 Maximum number of coughs in one hour in the six simulated scenarios, each 895 h of continuous cough monitoring

Simulation 1	Simulation 2	Simulation 3	Simulation 4	Simulation 5	Simulation 6
86	64	65	93	115	80

will have the tools they need to assess the effectiveness of their interventions accurately. Simply put, monitoring is the key to driving the science of cough forward, one data point at a time.

Author Contributions MR wrote the first version of the manuscript, performed the analysis, and prepared all figures, WJS collected the data, and PMS was coordinating the work. All authors contributed with updates to the first version of the manuscript and reviewed the final version.

Funding No funding was received to assist with the preparation of this manuscript.

Declarations

Conflict of interest MR and PMS are employees of Hyfe Inc., and WJS declares no related conflict of interest at the time of submitting this manuscript.

References

1. Cherry DK, Woodwell DA, Rechtsteiner EA (2007) National ambulatory medical care survey: 2005 summary. Adv Data 387:1–39 (**PMID: 17703793**)
2. An J, Lee JH, Won HK et al (2022) Cough presentation and cough-related healthcare utilization in tertiary care: analysis of routinely collected academic institutional database. Lung. https://doi.org/10.1007/s00408-022-00555-w (**PMID: 35810219**)
3. Gabaldón-Figueira JC, Keen E, Rudd M et al (2022) Longitudinal passive cough monitoring and its implications for detecting changes in clinical status. ERJ Open Res 8(2):00001–02022. https://doi.org/10.1183/23120541.00001-2022 (**PMID: 35586452;PMCID: PMC9108969**)

4. Woolf C, Rosenberg A (1964) Objective assessment of cough suppressants under clinical conditions using a tape recorder system. Thorax 19(2):125–30. https://doi.org/10.1136/thx.19.2.125 (**PMID: 14128569; PMCID: PMC1018810**)
5. Zimmer AJ, Ugarte-Gil C, Pathri R et al (2022) Making cough count in tuberculosis care. Commun Med 2:83. https://doi.org/10.1038/s43856-022-00149-w
6. Gabaldon-Figueira JC, Brew J, Doré DH et al (2021) Digital acoustic surveillance for early detection of respiratory disease outbreaks in Spain: a protocol for an observational study. BMJ Open 11(7):e051278. https://doi.org/10.1136/bmjopen-2021-051278 (**PMID: 34215614;PMCID: PMC8257291**)
7. Gabaldón-Figueira JC, Keen E, Giménez G et al (2022) Acoustic surveillance of cough for detecting respiratory disease using artificial intelligence. ERJ Open Res 8(2):00053–02022. https://doi.org/10.1183/23120541.00053-2022 (**PMID: 35651361;PMCID: PMC9149391**)
8. Kang YR, Oh JY, Lee JH et al (2022) Long-COVID severe refractory cough: discussion of a case with 6-week longitudinal cough characterization. Asia Pac Allergy 12(2):e19. https://doi.org/10.5415/apallergy.2022.12.e19 (**PMID: 35571551;PMCID: PMC9066079**)
9. Weisstein, Eric W. "Negative Binomial Distribution." From MathWorld--A Wolfram Web Resource. https://mathworld.wolfram.com/NegativeBinomialDistribution.html. Accessed 22 July 2022