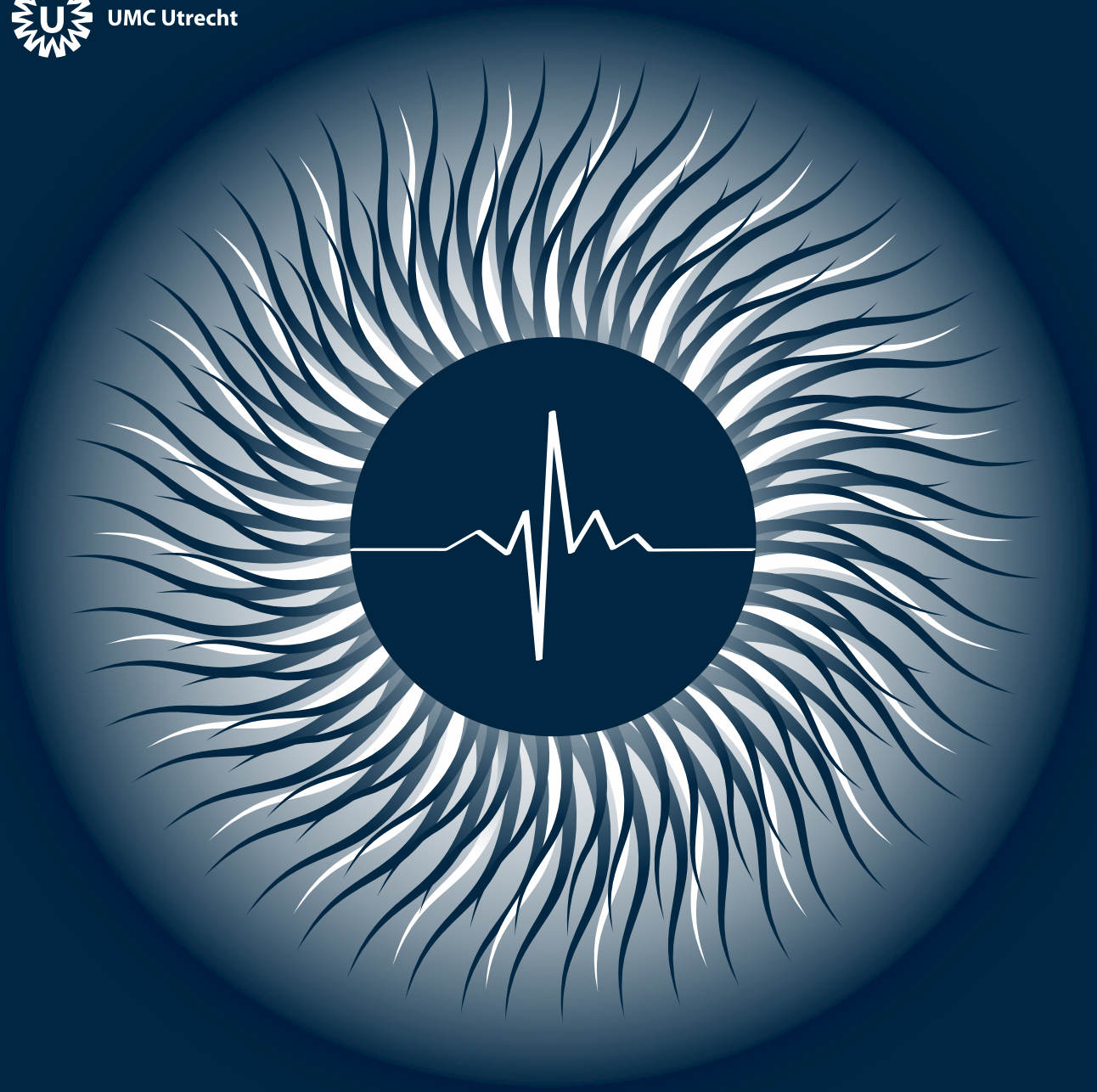
UMC Utrecht



Special issue

Vol 5 N°1

ISSN 2667-1204

# Journal of Trial and Error

## Scientific failure and uncertainty in the health domain

**Special Issue**

# Scientific Failure and Uncertainty in the Health Domain

Volume 5
Issue 1
[Date], 2025

ISSN 2667-1204
[DOI]



In cooperation with The New Utrecht School

Cover by Lieve Visser

Funded by the Social Innovation Program of the University Medical Center Utrecht

# Contents


**1-6** Editorial

The case for embracing trial-and-error in health research: Rethinking failure and uncertainty as foundations of progress
by *Stefan Gaillard, Stefan van Geelen, Elvire Landstra, & Arno Hoes*

**7-16** Editorial

How to fail successfully
by *Berent Prakken*

**17-23** Meta-Research

Issues in Clinical Studies Leading to Medical Research Ethics Committee (MREC) Negative Decisions
by *Sigrid E.M. Heinsbroek, Vincent Bontrop, Rutger P. Chorus, & Michel Zwaan*

**24-36** Empirical

The "Function" of Art?: Challeges of Setting Up Artistic Research Residencies in Elderly Care Institutions
by *Falk Hübner, Gjilke Keuning, & Marijke Lucas*

**37-48** Empirical

Medical Expert Endorsement Fails to Reduce Vaccine Hesitancy in U.K. Residents
by *Folco Panizza, Piero Rozani, Carlo Martini, Lucia Savadori, & Matteo Motterlini*

**49-57** Reflection

On the Significance of Place: Vaccination Refusal as a Situated Phenomenon
by *Martijn van der Meer*

**58-64** Reflection

Digital Nudges: A Reflection on Challenges and Improvements Inspired by the Gloria Adherence Subproject
by *David Grüning*

# Contents


**65-83**
Empirical

Cognitive Function, Mood and Sleep Quality after Two Months of Intermittent Fasting
by *Maja Batorek & Ivana Hromatko*

**84-90**
Reflection

Cognitive or Emotional Improvement through Intermittent Fasting? Reflections on Hype and Reality
by *Stephan Schleim*

**91-101**
Empirical

Smile, You're on Camera: Investigating the Relationship between Selfie Smiles and Distress
by *Monika Lind, Michelle Byrne, Sean Devine, & Nicholas Allen*

**102-106**
Reflection

A Smiling Paradox: Exploring the Constructed Nature of Emotions. A Reflection on the Relationship Between Smiling in Selfies and Distress
by *Anne Margit Reitsema, Sanne Nijhof, & Odilia Laceulle*

**107-121**
Empirical

Prenatal Sildenafil and Fetal-placental Programming in Human Pregnancies Complicated by Fetal Growth Restriction: A Retrospective Analysis
by *Fieke Terstappen, Torsten Plösch, Jorg J.A. Calis, Wessel Ganzevoort, Anouk Pels, Nina D. Paauw, Sanne J. Gordijn, Bas B. van Rijn, Michal Mokry, & A. Titia Lely*

**122-127**
Reflection

In the era of whole transcriptome sequencing: Reflections on the Molecular Genetic Effect of prenatal Sildenafil for Fetal Growth Restriction
by *Carsten F.J. Bakhuis & Marcel A.G. van der Heyden*

# Contents

**128-135**
Empirical

Partial Endothelial Trepanation versus Deep Anterior Lamellar Keratoplasty in keratoconus patients: Results of the PENTACON trial
by *Robert P.L. Wisse, Cathrien A. Eggink, Bart T.H. van Dooren, & Allegonda van der Lelij*

**136-140**
Reflection

Reflection on the PENTACON Trial: Lessons learned from an unpublished study
by *Robert P.L. Wisse*

# The case for embracing trial-and-error in health research: Rethinking failure and uncertainty as foundations of progress

Stefan Gaillard[1,2], Stefan van Geelen[3], Elvire Landstra[2,4], Arno Hoes[5]

[1]Institute for Science in Society, Radboud University Nijmegen, the Netherlands
[2]Center of Trial and Error, Utrecht, the Netherlands
[3]Education Center, University Medical Center, Utrecht, Netherlands
[4]Center of Research on Psychological disorders and Somatic diseases (CORPS), Department of Medical and Clinical Psychology, Tilburg University, The Netherlands
[5]University Medical Center Utrecht, Utrecht, the Netherlands

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
July 2, 2025
**Accepted**
July 9, 2025
**Published**
September 14, 2025
**Issued**
May 24, 2024

**Correspondence**
gaillard@trialanderror.org



**Keywords** *failure, uncertainty, health domain, trial-and-error*

Uncertainty is a daily reality in the health domain. Clinicians routinely face it when diagnosing patients, predicting outcomes, and selecting treatments in partnership with patients. Patients and their families, in turn, experience uncertainty about their condition and what the future may hold. Biomedical and clinical researchers aim to reduce this uncertainty through rigorous research, but not all of it succeeds. As others have argued, tolerating uncertainty may be one of medicine's next revolutions (Simpkin & Schwartzstein, 2016), yet the process of trial-and-error in science remains underacknowledged and underexplored. Sharing failures can save time and resources, ultimately improving patient outcomes. Still, most failed research disappears into the proverbial file drawer (Feng et al., 2024; Turner et al., 2022), leaving valuable lessons unlearned.

The international movement to "increase value and reduce waste" in health research has highlighted systemic inefficiencies at every stage of the research cycle – from question formulation to publication and implementation (Macleod et al., 2014). At the same time, translational medicine continues to face what is often called the "valley of death": the persistent gap between promising biomedical discoveries and their successful application in clinical practice (Seyhan, 2019). What emerges from the contributions in this special issue is that failure in health research is not always due to simple incompetence or lack of rigor, but rather often the result of complex, context-dependent dynamics: interventions that worked elsewhere fall flat in new cultural or institutional settings; sound experimental designs encounter logistical or ethical barriers; promising innovations reveal unanticipated side effects, diminishing returns, or practical incompatibilities when scaled. We therefore believe that one underrecognized source of waste is the neglect of *rigorous failure*: research that is methodologically and logistically well-designed and well-justified, but does not lead to successful outcomes. This includes, on the one hand, studies that yield null or negative results — such as the absence of an expected effect — and, on the other hand, well-reasoned methodological or logistical innovations that ultimately prove unworkable in practice. Both types of failure can offer crucial insights for researchers facing similar uncertainties yet are rarely shared or systematically analyzed. This opacity surrounding failure is driven not only by cultural discomfort, but also by institutional disincentives — funding structures which leave no room to publish failure, journals which reject failed research more often than not, and tenure tracks which penalize failure.

A pattern across the studies published in this issue is that failure often emerges at the

point of translation: where evidence, ideas, or technologies must move between different domains — between lab and clinic, between cultures, between trial protocol and real-world practice. This translational friction reveals that success in one context does not guarantee generalizability. Another recurring insight is that *failure often exposes implicit assumptions* — about how people behave, how systems function, or how knowledge travels — that only become visible when things do not work as expected. These productive failures are especially instructive, as they force researchers to recalibrate their models, expectations, and designs. Importantly, these productive failures are often generalizable to other researchers attempting similar translations. By reflecting on both small setbacks and larger failures, we aim to foster a culture of continuous improvement that enhances scientific progress throughout the research cycle. Through the contributions in this issue, we highlight how embracing failure as a constructive and integral part of research can lead to more innovative solutions and thus better patient outcomes and improved health.

Publishing failed studies offers several key advantages that can significantly improve the efficiency and quality of health research. One of the most direct benefits is the prevention of duplicate work. When failed studies are shared, researchers can learn from one another, ensuring that they do not repeat the same unsuccessful lines of inquiry. This not only saves time but also prevents the waste of valuable funding. Avoiding redundant efforts is thus crucial for advancing patient care.

Another important advantage is the ability to gain insights into how things *do* work by learning from the failures of others. Often, failures contain valuable lessons that other studies may overlook. By understanding why certain approaches did not work, researchers can refine their hypotheses, improve their study designs, and ultimately avoid pitfalls. This process of learning from failure accelerates innovation and fosters a deeper understanding of complex systems, especially in scientific disciplines where the path to success is not always straightforward.

Moreover, sharing negative or null results contributes to more robust meta-analyses. Scientific knowledge is rarely derived from a single study; it typically emerges from the accumulation and synthesis of data across multiple, rigorously conducted studies. When negative or null results from rigorous studies are made available, they become part of this broader analysis, offering a more complete picture of the evidence. Meta-analyses that include all types of results provide a clearer, more reliable view of a research question, allowing for more accurate conclusions and better-informed decisions in healthcare.

Finally, the increasing employment of machine learning in research benefits from the publication of failed studies. Machine learning algorithms rely heavily on large, diverse datasets to train and optimize models. When failures are included in these datasets, they enhance the quality of the machine learning output by providing the necessary variety of outcomes that can lead to more accurate predictions and solutions. By incorporating both successes and failures, machine learning models can become more robust and balanced, and thus more useful for research and clinical applications.

## Institutional context

This special issue is a collaborative effort between the Journal of Trial and Error and *The New Utrecht School*, sponsored by the Social Innovation Program at the University Medical Center (UMC) Utrecht. While the conceptualization of this issue took place in Utrecht, the Netherlands, contributions span a broad range of research institutions from multiple countries, reflecting the global relevance of failure and uncertainty in health research. These contributions highlight the diverse ways in which failure and uncertainty manifest across various health disciplines, settings, and research contexts. The contextual nature of failure and uncertainty means that their impact can vary greatly depending on the specific circumstances, making it essential to reflect on these variations. For this reason, we have paired empirical articles with reflection pieces to foster a deeper understanding of how failure can be approached and learned from across different contexts.

Parallel to the publication process, we organized a series of monthly lunch lectures focused on scientific failure and uncertainty

in the health domain. These lectures served as an important complement to the Adrienne Cullen Lecture, which already exists at the UMC Utrecht as a prominent forum for discussing failures in clinical practice — but not for discussing failure in biomedical research. The lunch lectures provided an informal yet intellectual space for researchers, clinicians, and students to engage in discussions about how failure and uncertainty are addressed in health research and practice. These conversations fostered a deeper appreciation of the role of failure in the scientific process and encouraged participants to reflect on the potential benefits of embracing failure on research outcomes, clinical practices, and health outcomes.

## Contributions

*How to fail successfully* by Berent Prakken offers a personal and reflective account of failure in translational medicine, centered on a collapsed biotech collaboration around peptide immunotherapy for arthritis. Through vivid narrative and contextual insight, Prakken argues that failure should be understood as an instance of knowledge creation. Drawing from clinical, entrepreneurial, and educational experiences, he advocates for training translational scientists to embrace uncertainty, complexity, and setbacks as part of their professional identity and ethical responsibility.

*Issues in clinical studies leading to Medical Research Ethics Committee (MREC) negative decisions* by Heinsbroek et al. investigates the reasons behind negative decisions issued by MREC NedMec on clinical research proposals over a 5-year span. The study identifies frequent shortcomings such as incomplete research files, weak or missing scientific rationale, inadequate benefit-risk analyses, and insufficient or unclear consent procedures. These issues often lead to rejections, but many of these problems are preventable by submitting well-prepared, compliant proposals. The authors recommend that researchers consult with ethicists, methodologists, or regulatory bodies early in the design process to improve the likelihood of approval and to uphold ethical standards.

In *The "function" of art?: Challenges of setting up artistic research residencies in elderly care institutions*, Hübner et al. examine three 6-month artistic residencies in a Dutch elderly care home, where artists explored how art could address relational shifts during residents' transition to institutional care. The project faced challenges, including staff expectations, reliance on language over art, and lack of creative space. Despite these tensions, the authors argue that art's "functionless" nature allows for open-ended engagement, shifting how healthcare institutions view meaning, connection, and care. They call for more thoughtful project design, earlier engagement with art, and dedicated spaces for creative practices in social care settings.

A large, preregistered longitudinal experiment conducted by Panizza et al. during the 2021 COVID-19 UK vaccination campaign, *Medical expert endorsement fails to reduce vaccine hesitancy in U.K. residents,* contradicts findings from a similar Italian study. Participants received messages correcting vaccine misconceptions, each endorsed by medical experts. Contrary to prior findings from a similar Italian study, the expert-backed messages had no significant effect on participants' vaccination intentions, beliefs, or uptake. The authors discuss possible explanations, including timing of the UK campaign, cultural context, and ceiling effects due to already high baseline vaccine support. They conclude that expert endorsement alone may not be a universally effective strategy and emphasize the need for more context-sensitive interventions.

In *On the significance of place: Vaccination refusal as a situated phenomenon,* Martijn van der Meer continues and deepens this reflection. Interpreting the UK trial as a conceptual replication, Van der Meer argues that expert messaging is not universally effective and critiques the "knowledge deficit model" that assumes hesitancy stems from ignorance. Drawing from medical humanities and anthropology, he proposes that trust, local context, and social dynamics shape how vaccine refusal operates. Public health campaigns, he concludes, should adopt more dialogical, participatory strategies tailored to specific communities instead of assuming one-size-fits-all scientific messaging will succeed.

David Grüning's article *Digital nudges: A reflection on challenges and improvements inspired by the Gloria Adherence Subproject* offers a critical reflection on digital nudging in-

terventions, using the Gloria Adherence Subproject as a case study. The original project, reported in an earlier issue of the Journal of Trial and Error, aimed to improve medication adherence via a daily app reminder yet showed no significant effect. Grüning identifies three broader challenges of digital nudges: they can undermine user agency, their effects decay quickly over time, and they often fail in complex ("wicked") environments. He proposes "boosting" as a more sustainable alternative — interventions that build user competencies. Grüning recommends shifting from simple reminders to empowering participants with clear information and adaptive tools.

Tackling a popular "biohacking" idea, Batorek and Hromatko examine the cognitive impacts of a 2-month time-restricted eating (TRE) regimen in *Cognitive functions, mood and sleep quality after two months of intermittent fasting*. The study tested the effects of TRE on cognitive performance, mood, and sleep in healthy adults. Although the experimental group lost weight, no statistically significant differences were found between the fasting and control groups on a battery of cognitive tasks or self-reported measures of mood and sleep quality. Improvements observed across both groups were attributed to practice effects or seasonal mood changes. The authors urge caution against overhyping the cognitive and psychological benefits of intermittent fasting and highlight the need for better-controlled, randomized studies before endorsing TRE as a cognitive or mental health intervention.

In his commentary *Cognitive or emotional improvement through intermittent fasting? Reflections on hype and reality*, Stephan Schleim reviews the previous study, highlighting methodological limitations such as the short intervention duration, lack of randomization, and limited ecological validity of cognitive tests. He contextualizes the findings within broader debates on neuroenhancement and cautions against inflated expectations, noting that even pharmacological interventions show only modest cognitive effects in healthy individuals. Schleim concludes that while intermittent fasting remains a topic of public and scientific interest, reliable evidence for its mental benefits remains lacking and nuanced, long-term research is needed.

The exploratory pilot study *Smile, you're on camera: Investigating the relationship between selfie smiles and distress* by Lind et al. investigated whether smiling behavior captured in selfie videos via a mobile sensing app (EARS) reflected self-reported psychological distress in college students. Contrary to the hypothesis based on Paul Ekman's expressed emotion framework — that smiling decreases as distress increases — the study found that smiling intensity either stayed the same or even increased with higher levels of stress, anxiety, and depression. These counterintuitive results align more with Alan Fridlund's behavioral ecology view, which posits that facial expressions serve communicative, rather than expressive, functions. The study highlights limitations such as small sample size, lack of positive affect measures, and simplistic automated smile detection, but nonetheless contributes to debates about the validity of facial expressions as emotional indicators. The findings suggest that smiling may serve social or self-regulatory functions under stress and demonstrate the potential and the pitfalls of mobile sensing and automated facial analysis in psychological research.

Reitsema et al. provide a reflective commentary on this work, interpreting its counterintuitive finding through the lens of constructed emotion in *A smiling paradox: Exploring the constructed nature of emotions. A reflection on the relationship between smiling in selfies and distress.* Challenging Paul Ekman's idea of universal, biologically fixed emotional expressions, the authors align with Lisa Feldman Barrett's view that emotions are context-dependent and shaped by past experiences, social norms, and individual goals. They argue that the act of recording oneself likely heightens self-awareness and social pressure to appear positive, prompting smiles that do not reflect genuine emotional states. The commentary also critiques the oversimplified assumptions behind AI-powered emotion recognition systems, warning against the ethical and practical risks of interpreting facial expressions as direct indicators of emotion. Ultimately, Reitsema et al. call for more nuanced, context-sensitive methods in both research and technology to capture the true complexity of human emotional life.

In *Prenatal sildenafil and fetal-placental programming in human pregnancies complicated by fetal growth restriction: A retrospective gene*

*expression analysis*, Terstappen et al. describe their study which investigated the impact of prenatal sildenafil treatment on placental and fetal gene expression in pregnancies affected by fetal growth restriction (FGR). Despite the STRIDER trial showing no clinical benefit from sildenafil in FGR, this study reveals that sildenafil does exert biological effects on gene expression relevant to fetal development and long-term cardiovascular and renal programming. Specifically, the authors used gene expression and enrichment analyses to identify changes in pathways related to smooth muscle proliferation, DNA repair, vascular development, and kidney function.

The accompanying reflection focuses on the value of studies that capture the often subtle and complex molecular effects that can occur even when no gross differences are observed. In their piece *In the era of whole transcriptome sequencing: Reflections on the molecular genetic effect of prenatal sildenafil for fetal growth restriction*, Bakhuis and Van der Heyden emphasize these effects on placental and fetal gene expression. Their reflection highlights the importance of such mechanistic studies, recommends broader analytic approaches (e.g., epigenetics, proteomics), and advocates for integrating molecular profiling into all clinical trials — especially when repurposing drugs — to better understand their multifaceted effects and to inform safer, more effective treatments for FGR.

In *Partial Endothelial Trepanation versus Deep Anterior Lamellar Keratoplasty in keratoconus patients: Results of the PENTACON trial*, Wisse et al. document the trial-and-error process of the PENTACON trial, a multicenter randomized study comparing Deep Anterior Lamellar Keratoplasty (DALK) with the newer Partial Endothelial Trepanation (PET) technique in keratoconus patients. The trial was prematurely terminated due to under-enrollment, evolving clinical practices, and loss of equipoise. Despite technical promise, both techniques showed unexpectedly high complication rates and a steep learning curve, with no clear safety advantage for PET over DALK. Although underpowered, the study illustrates critical challenges in surgical trials — balancing methodological rigor, innovation, and patient safety — and highlights the ethical responsibility to report inconclusive or halted trials to advance evidence-based practice.

In *Reflection on the PENTACON Trial: Lessons learned from an unpublished study*, Robert Wisse candidly reflects on the failure of the PENTACON clinical trial, which was terminated due to poor recruitment, protocol rigidity, and loss of clinical equipoise. Beyond the halted trial, Wisse draws valuable lessons about trial design, feasibility, mentorship, and the ethical imperative to publish negative results. His experience underscores the need for adaptive protocols, stronger support for early-career researchers, and structural reforms to mitigate publication bias. While scientifically inconclusive, the PENTACON trial became a powerful teacher, shaping Wisse's more resilient and ethically driven research ethos.

## Conclusion

Although the focus of this special issue is on health science, the taboo surrounding failure is a broader cultural problem. Changing the culture surrounding failure will be a difficult endeavor, as cultural change often is. One crucial component for cultural change is education. By fostering an environment in which failure is recognized not as a setback but as an opportunity for learning, we can reshape how both researchers and the wider public perceive scientific progress. Encouraging open dialogue about failures, incorporating failure as a learning tool in academic curricula, and supporting transparency in research will be essential steps in shifting this cultural norm.

Ultimately, embracing failure in science can catalyze more robust, innovative solutions and improve the translation of research into real-world benefits. As the studies in this issue demonstrate, rigorous failures, when shared and understood, offer valuable insights that can accelerate progress and improve patient outcomes. The path forward will require effort, honesty and commitment, but the increased value — a greater transparency, more effective research, and better-informed health decisions and outcomes — are well worth the challenge.

## References

Feng, Q., Mol, B. W., Ioannidis, J. P., & Li, W.

(2024). Statistical significance and publication reporting bias in abstracts of reproductive medicine studies. *Human Reproduction, 39*(3), 548-558. http://doi.org/10.1093/humrep/dead248

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P., Al-Shahi Salman, R., Chan, A., & Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *The Lancet*, *383*(9912), 101-104. http://doi.org/10.1016/S0140-6736(13)62329-6

Seyhan, A. A. (2019). Lost in translation: The valley of death across preclinical and clinical divide–identification of problems and overcoming obstacles. *Translational Medicine Communications*, *4*(1), 1-19. http://doi.org/10.1186/s41231-019-0050-7

Simpkin, A., & Schwartzstein, R. (2016). Tolerating uncertainty — the next medical revolution? *New England Journal of Medicine*, *375*(18), 1713-1715. http://doi.org/10.1056/NEJMp1606402

Turner, E. H., Cipriani, A., Furukawa, T. A., Salanti, G., & de Vries, Y. A. (2022). Selective publication of antidepressant trials and its influence on apparent efficacy: Updated comparisons and meta-analyses of newer versus older trials. *PLOS Medicine*, *19*(1), Article e1003886. http://doi.org/10.1371/journal.pmed.1003886

**Correction notice**
Incorrect Special Issue Labeling (Article erroneously excluded): This article was previously not labeled as part of a special issue due to an error. This has now been corrected.

# How to fail successfully

Berent Prakken[1]

**Keywords** *translational medicine, immune tolerance, education innovation, translational scientist*

> *"Boys we were - but good boys. If I may say so myself. We're much smarter now, so smart it's pathetic." (Nescio, 2012)*

This quote from the Dutch writer Nescio (1933/2012, p.35) perfectly captured how I felt that morning in Paris in 2005. Salvo Albani and I were sitting in a coffee bar — the kind that actually sells coffee. Outside, the rain was falling steadily, while inside, it was warm and humid. The coffee tasted bitter.

I had arrived in Paris the day before, traveling by train from Utrecht, in high spirits. Salvo, who had flown in from San Diego, was equally upbeat. He was on the verge of signing an agreement with a venture capital firm that evening, a deal that would secure funding for his biotech spin-off. This funding was crucial — it would finance the next, pivotal clinical trials that we hoped would definitively prove the effectiveness of peptide immunotherapy for arthritis. Less than 24 hours later, everything had changed. It was so close — so terribly close. And then, it was out of reach again.

## First encounter

Salvo and I had first met eleven years earlier, in 1994, in Pavia. The occasion was the European Pediatric Rheumatology Congress, where I presented two abstracts to an international audience for the first time. I had meticulously prepared for the meeting, anticipating the kinds of questions the sharpest minds in the audience might pose. I even prepared two additional slides as a reserve to address the most challenging questions I could imagine. Twice, a tall, well-dressed and good-looking young Italian man with a distinctive accent approached the microphone and asked exactly the two questions I had prepared for. It was Salvo. After my second presentation, he approached me, introduced himself, and said: "Since you had slides ready to answer my questions, everyone in the audience now thinks we're secretly working together. So why don't we start working together now?" That marked the beginning of a close collaboration and a friendship that has lasted for nearly 30 years. During these years we did not reach our target (a brilliant breakthrough treatment for autoimmunity), but while failing to do so we accidentally achieved something with perhaps more long-term impact: the establishment of an institute for translational scientists — Eureka!

On the surface, we seemed as different as our birthplaces: Amsterdam (mine) and Siracusa (Salvo's). Salvo was more outspoken and bolder; I tended to be quieter and more cautious. Yet, we soon discovered we had far more in common than what set us apart. We were both pediatric immunologists and deeply passionate about translational science, that is, science with true impact. After our first meeting, our professional relationship quickly deepened, ultimately even evolving into a close friendship. We shared our struggles with journal submissions, horrific reviewers, and the challenges of managing a growing lab. For me, the most important thing was probably the reassurance of knowing that someone else, albeit on the other side of the world, was thinking along the same lines.

Looking back, I can see clearly that the success of our personal partnership was built on three key elements: mutual respect, trust, and a shared passion to make a difference.

## The journey starts

At the time of our first meeting I was doing my

[1] UMC Utrecht, Utrecht, the Netherlands

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
October 16, 2024
**Accepted**
March 3, 2025
**Published**
April 1, 2025

**Correspondence**
UMC Utrecht
bprakken@umcutrecht.nl

PhD at the lab of professor Willem van Eden, at the time world-famous for his discovery of a link between heat shock proteins and arthritis (van Eden et al., 1988). Salvo already had his own research group at the University of California San Diego and had recently made a pivotal discovery on the role of molecular mimicry in arthritis that also involved a peptide derived from a heat shock protein (Albani et al., 1995). Following our first meeting in Pavia, things developed rapidly. The more we talked the more we realized in how many ways we were alike. After finishing my PhD in Utrecht I moved my family to San Diego and joined Salvo's group for the next two years. By the time we met in Paris in 2005, Salvo was working at the world-renowned Burnham Institute for Molecular Medicine in San Diego, while I had returned to the Netherlands and built my own lab at the University Medical Center Utrecht. Our two labs collaborated closely: We shared data, patient samples, antibodies, and even original ideas. The researchers in our two groups also felt the connection, which was further improved with the exchange of personnel.

We were relatively young – in our early forties, by now typical early-mid career professionals. The same career phase during which today, unfortunately, many young scientists often decide to leave the life sciences field and seek a career elsewhere. Salvo and I never felt the urge to leave science. Looking back, I see two reasons for our resilience. Firstly we were in this together and, secondly, we saw the large unmet need in the patients we were treating every day. Quitting was simply not an option.

## The unmet need

Salvo and I were both pediatric immunologists, treating children with Juvenile Idiopathic Arthritis (JIA) in our clinics. In 1994, when Salvo and I met, the only effective treatment for JIA was Methotrexate (MTX), a drug developed decades earlier primarily for other conditions, such as malignancies (Calasan & Wulffraat, 2014).

MTX was and in many ways still is a mysterious drug. It was initially proposed for the treatment of autoimmune diseases because of its immunosuppressive effects. However, based on the dosage used, it was highly unlikely that it had any significant immunosuppressive impact. Indeed, over the years, no clear immunosuppressive effect has been confirmed, either *in vitro* or *in vivo*. While MTX was proven effective, it had clear limitations. The most prominent was the severe nausea it induced in the majority of patients (Calasan et al., 2013). This nausea could become so intense that children would feel sick merely at the thought of taking MTX, upon hearing the word "MTX," or even when seeing the same yellow color that the pills have. Another major issue was that, although MTX suppressed the symptoms, in many cases it did not provide a lasting solution: a disease- and medication-free remission. In other words, it did not cure the disease.

## Time travel

Until now, treatments for autoimmune diseases have primarily focused on suppressing symptoms. Over the years, we have made tremendous strides in symptom management, often with life-saving outcomes. Consider the discovery and use of insulin for Type 1 Diabetes. This was certainly a groundbreaking treatment (Lewis & Brubaker, 2021), but today, almost 100 years after Sir Frederick Banting's discovery of insulin at the University of Toronto, diabetes patients still mostly rely on insulin to manage their symptoms. We felt the current approach was taking us in the wrong direction. What we envisioned was something radically different: a disease- and medication-free remission—a true cure. In essence, we wanted to travel back in time to the period before the patient became ill. At the time, this was science fiction, and sadly, it largely remains so today.

## Under fire

Our vision was to reprogram the gatekeepers of inflammation—CD4+ T cells—into T cells capable of modulating inflammation. It would make sense that such cells existed simply based on the principle: What goes up, must come down. And we believed, based on data mostly obtained in animal models, that this should be possible in a so-called antigen specific manner (Prakken et al., 2002). By using antigen-specific immune modulation we thought we could bypass systemic side effects and specifically control the immune response

(Albani et al., 2011; Albani & Prakken, 2006), an approach Salvo likened to a dimmer. However, not everyone agreed with us. And that was a problem.

Our critics, mostly anonymous reviewers with probably far more seniority than we had at that time, were quick to point out why we all got it wrong. First, we began our work on T cells with immune regulatory capacity while just a few years earlier, another type of "suppressor" T cells (CD8+ T cells) had fallen out of favor in the scientific community after a brief period of popularity. Although we were targeting a different subset of T cells (CD4+ T cells) under a completely different hypothesis and framework, the mere association with "suppressor" T cells was enough to reject our work. At that time, having the term "suppressor T cells" in the title of a paper almost guaranteed rejection. But this was only the beginning of the challenges we faced.

Second, we were attempting to modulate T cells using mucosal tolerance induction via nasal application of peptides. The problem was that about a decade earlier, other researchers had introduced oral tolerance induction for autoimmune diseases, creating a sensation in the field (Trentham et al., 1993; Weiner et al., 1994). This led to a surge of excitement and numerous papers on mucosal tolerance, particularly in autoimmune diseases. This enthusiasm culminated prematurely in clinical trials using oral myelin basic protein for MS and oral collagen for Rheumatoid Arthritis, both of which unfortunately failed. Despite significant differences between this and our approach, the failure of these trials led to widespread dismissal of the field by the scientific community.

Third, we applied peptides derived from heat shock proteins (HSPs). HSPs are evolutionarily conserved proteins that are also immunodominant. Since HSPs are upregulated during cell stress, they were considered potential candidates for inducing antigenic mimicry, possibly leading to autoimmunity. However, attempts to use recombinant proteins for immune modulation had shown that some of the remarkable positive effects observed in experimental models were due, at least in part, to contamination with bacterial compounds. Although we used peptides instead of proteins and thus avoided this issue, the reputation of HSPs was already tarnished—they also had fallen out of fashion.

So, we were using — and even combining — three different methods, each of which was unpopular, to say the least. Some of our mentors advised us to pursue a less contentious direction, but we were too convinced of our approach to follow their advice. Perhaps it was naïveté or even arrogance, but we believed we had strong intellectual arguments for choosing this path. Our reviewers and competitors, however, were far less convinced of the soundness of our choices, and they were not shy about making their opinions clear. One review I received, which I've meticulously saved over the years, ended with the line: "Altogether, this can be regarded as an interesting yet completely futile experimental exercise."

Why didn't we do the smart thing and abandon our risky approach? In a way, instead of discouraging us, these negative reviews almost had the opposite effect: Since we saw little intellectual merit in the often-harsh critiques, our confidence only grew. And there was another factor: We were not alone. In hindsight, I realize that we were practicing team science before it became a recognized and fashionable concept — which, fortunately, it is becoming today.

## More poor timing

If you think our timing was poor, that wasn't the half of it. We embarked on this work at a time when MTX was still the only proven drug for severe cases of autoimmune diseases. We were unaware that a new class of drugs was about to revolutionize the field of rheumatology: biological agents that block cytokine pathways. Beginning with anti-TNF-alpha therapy, these drugs had a profound impact, demonstrated first in Rheumatoid Arthritis, then in Juvenile Idiopathic Arthritis, and in many other autoimmune diseases. Their efficacy in suppressing symptoms was impressive. However, in our view, this did not diminish the need for the specific tolerance-inducing treatments we had in mind. These cytokine and pathway-blocking therapies were non-specific and therefore carried the potential for side effects. Although short-term side effects seemed mild, early treatments for example revealed the re-emergence of latent tuberculosis. More-

over, these therapies did not cure the disease: Symptoms almost invariably returned once treatment was stopped. This essentially meant that patients would require lifelong treatment, thereby increasing the risks of long-term side effects. Blocking a cytokine pathway for a short period may be relatively safe, but doing so for years, or even decades, could significantly increase risks, such as vulnerability to microbial agents or disruption of intrinsic immune homeostasis. We were convinced that there remained a critical need to develop alternative therapies to achieve our ultimate goal of "time travel": restoring the natural immune homeostasis that existed before the onset of autoimmune disease.

## The lab in which the sun never sets

The playing field had changed, and the bar was now set even higher than before. But instead of being discouraged, we continued on the same track working nonstop to explore new ways to induce immune tolerance. Our two labs collaborated closely: We shared data, patient samples, antibodies, and even a sense of camaraderie. The two groups were very complementary and despite the cultural and organizational differences, we created a strong team with unifying traditions and habits and a thriving intrinsic motivation – now arising not only from just us two but also (and most importantly) from other lab group members. We referred to our partnership as "the lab in which the sun never sets", because with the time difference, someone was always conducting an experiment somewhere. Both lab entrances were adorned with a name plate reading "Iacopo Institute for Translational Medicine"; named for Salvo's giant friendly dog who would always accompany us on our walks in San Diego. At that time translational medicine was not yet a buzz word, which in subsequent years it has become.

We identified suitable peptide epitopes and explored different ways to administer those peptides in various experimental models, including combinations with stem cell therapy and blockade of the TNF-alpha pathway (Delemarre et al., 2011; Delemarre et al., 2014; Kamphuis et al., 2005; Kamphuis et al., 2006; Zonneveld-Huijssoon et al., 2012). Looking back, I—naively—placed far too much trust in the significance of animal models. I was captivated by the simplicity and endless possibilities of these models. It took time even for me, a clinician at heart, to fully appreciate the vast gap between model systems and real-life clinical settings in actual patients. Looking back this may be one of my biggest failures.

My lab also studied specific immune responses in patients during different stages of disease, looking for patterns that may give us clues about the natural immune regulatory response. These studies benefited from the excellent clinical organization in Utrecht set up by Nico Wulffraat and colleagues, which allowed us to do precise testing in well-defined cohorts of patients. To enhance the immune studies in patients we also developed new techniques, such as the multiplex immune assay for use in human samples. For this we worked together with Vicki Seyfert as a core facility of the Immune Tolerance Network of NIH (de Jager et al., 2005). Salvo on the other hand did more basic immunology studies and took the most crucial step towards application in patients. He pulled off something truly amazing: a successful Phase I/II clinical trial with dnaJP1, the peptide developed in his lab (Koffeman et al., 2009; Prakken et al., 2004). At the time, his study design was quite unique; not just looking for side effects but also using extensive immunological testing of patient samples for indications of successful tolerance induction *in vitro* in patients using surrogate parameters in peripheral blood samples. It turned out that this therapy was not only safe, but there also seemed to be a stunning correlation between the observed immune tolerance induction in patient cells and clinical outcomes (Prakken et al., 2004). The resulting publications led to renewed attention to T-cell specific immune therapy. Salvo also did something else I did not even consider: He launched a company to secure the funding and organization needed for a large multicenter placebo-controlled study in patients with RA. I admired (and still admire) Salvo for his boldness and his persistence in making this happen. I never seriously considered this option — I probably am too much of an academic, and certainly too cautious to take this step.

Besides, I cherished my academic freedom and loved working with the technicians, PhD and Master's students, and postdocs in my

growing lab. I now realize that I had this luxury of focusing on academic lab work because of Salvo's incessant search for funding and strong connections with potential investors.

## Back to Paris

Salvo was in Paris to do exactly that, talk with investors. I joined him as I usually would when he was close to Utrecht. It allowed us to talk freely and coordinate our plans. We often thought along similar lines and moved quickly, with the risk that we ended up doing exactly the same thing. Thus, it was important we talked regularly. In these years long before COVID, Zoom was unavailable and Skype barely operational, so apart from weekly calls on regular phone lines, we used every opportunity to meet. Over the years we would meet in Rome, Chicago, Utrecht, Tucson, London, Lyon, New York, Berlin, Singapore, Genoa and many more places. This time, as Salvo was in Paris, I took the train to meet him and this was why we were now sitting across from each other in this dull cafeteria the morning after his meeting with prospective investors. The next studies, which would provide final proof of his approach, could not be achieved as the investors unexpectedly pulled out, at least for now. We were now sitting, disappointed, in this dreary bar in gray rainy Paris.

## A Eureka moment

Our meetings would often not only be about immunology and all our shared lab projects, but also venture into other subjects –our families, relationships (mine being significantly more monotonous than Salvo's), books, and life in general. These areas would, weirdly enough, never include politics. Primarily because we did not like the opportunistic and simplistic nature of politics and especially politicians. But we also realized that arguing about political issues would not get us anywhere; an advice by the way that I would like to share with anyone.

I do not recall who brought up the topic – as there often was strong synchronicity in our thinking, one of us probably started to vocalize what the other was thinking. But at some moment we found ourselves talking about what a difference it could have made if we knew ten years ago what we knew now. And from there it was a small step towards our Eureka moment: Why not let other, younger colleagues benefit from our past (and present) failures and thus prevent them from making the same mistake?. The atmosphere changed from gloom to excitement. It felt as if a new window was opened, blowing in fresh air. And above all, it opened a totally new field in which we frankly did not have much knowledge: How can translational scientists learn from our failures? How do people learn anyway? Often crazy ideas tend to stay at that very first stage — just an idea. Not for us, when we would collaborate. When an idea struck, giving up simply became unthinkable. From that moment on we kept at it, brainstorming, searching for people with similar mindsets and talking to them about our ideas. Most of the people we talked to considered our plans crazy and unrealistic but a small group of people was very enthusiastic and encouraged us to go on.

The plan gradually evolved into an Institute for Translational Medicine focusing on training young translational scientists and building a network for translational research. Part of the plan was to make Siracusa (Sicily), the birth town of Salvo, the homebase of the new institute because of its many possibilities, and the benefit of a beautiful historical city with an impressive cultural and scientific past. And thus, precisely a year later, on a warm Summer day in 2006, we traveled to Siracusa to set up the first legal entity that would be the start of the institute. After signing the paperwork in a hot notary's office full of books, binders, and papers, we went for a swim in the Mediterranean Sea. I will never forget the feeling as we dived into the water — it felt as if we were going to change the world.

## From Eureka moment to Eureka Institute

The problem was that we barely had more than a broad outline of an idea. Or maybe that was our strength: We did not foresee all the problems that could arise if we followed up on this idea. Being ignorant about the potential risks and difficulties helped us to simply go ahead and move, step by step, towards our goal.

That Summer I spent 2 months in San Diego for a short sabbatical with Salvo and we continuously kept thinking and talking about it —

sometimes we even forgot to talk about immunology. We slowly but steadily worked on a plan that with every meeting became more and more concrete.

There was no way back anymore. A year later I found myself together with Salvo again in the same hot notary's office in Siracusa, to officially launch the institute that we now decided to call Eureka. And again one year later, in 2008, we were ready for a huge next step: We had secured some minor grant funding. It was just enough to invite a small group of opinion leaders in the field of translational medicine, who we hoped were likeminded, to join us for an inaugural meeting of this new institute in Siracusa. I was pretty nervous about this first meeting. What if they just did not get it? What if they thought that there was no problem at all? What if they understood the problem but did not like our solution?

The meeting was held in a small room, almost without daylight, in the basement of a hotel in Siracusa. The group included Janet and David Hafler from Yale, Norm Rosenblum from the University of Toronto, Lucca Guidotti (then UCSD, now San Raffaela), Juan Carlos Lopez (Nature Medicine) and others; all experts with very different backgrounds but all committed to Translational Medicine. My worries were unnecessary: It was one of the most inspiring meetings I ever had. The team quickly found a joint mission: The institute should inspire, mentor, and educate translational researchers worldwide, bring them together in a network, and help to create impactful research. From there on the developments went quickly. Already at the same meeting a plan was drafted for an international certificate course to be held in Siracusa one year later.

## Into reality

The first Eureka course was held in Siracusa in 2009 with young translational scientists from all over the world participating. The course followed a unique blend of knowledge of the translational medicine pathway offered by experts in the field and educational innovation. The latter was mainly provided by Janet Hafler, a prominent medical educator from Yale. Thanks to her input, the course structure was rock-solid and different from what both participants and faculty were used to: It was truly learner-centered and set in an environment that encouraged interaction, discussion, and self-reflection. In that setting the participants met with peers who came from different countries and institutions with different cultures, but who were struggling with similar issues. Just listening to each other's experiences and realizing that they were not alone impacted the participants. The presence of renowned international faculty members with abundant experience and knowledge on every aspect of the translational medicine pathway was the other factor. In a conventional model, faculty staff would be teaching the participants by giving impressive lectures about their own achievements. Their role at Eureka was different: They were in the course to listen to the participants, support, and advise them. And in their presentations, they were asked not to just simply speak about their successes but also about their own insecurities and yes, about their failures.

## Life and career changing?

Already at this first try it became clear that we had started something special. Though still in its first raw version, and thus with inconsistencies and shortcomings, the course turned out to have a major impact on both the participants and the faculty. Something magical happened — already during the very first edition of the course participants were saying that the course was life and career changing. But what does "life and career changing" mean? Now, 15 years later I still do not completely understand this. Clearly there is something that Salvo many years later called "the two souls of Eureka". One soul being the access to expertise and knowledge of the full translational pathway – you cannot play a game if you do not know the rules of the game. And how better to learn these rules than from true experts in the field, the ones who had done it, who had gotten their hands dirty and understood it not just from theory but from practice? The other soul could be the deliberate exposure to complementary or 21$^{st}$ century skills such as critical thinking, problem solving, communication, and creativity. And then, maybe a third soul is the hidden curriculum: a safe environment in which the participants can freely interact with the faculty in such a way that the classi-

cal division based on seniority and hierarchy becomes irrelevant and even non-existing. In peer mentoring sessions they learn from each other's problems which often bear remarkable similarities to their own issues and worries. It strengthens the feeling that you are not alone in the so-called Valley of Death between bench and bedside. Last and not least it helped to bring people back to the reason why they once started to do research, namely to search for discoveries that really can make a difference for patients. With the entrance of role models such as Pat Furlong – parent and advocate for patients with muscular dystrophy – in one of the first courses, patients and societal impact literally came to stand in the middle of Eureka.

Young translational scientists grow up in a hypercompetitive system that forces them to focus on grants and papers to be successful, while they lose their strong intrinsic motivation. At least this is how it feels for them. Paradoxically over the years we could see that when they let go of this extrinsic motivation and focus on the "why" of their research, classical success in the form of papers and grants followed also.

It was a lucky coincidence that the growth of Eureka coincided with the international movement towards Open Science and the development of a new system for recognition and reward of scientists (Benedictus et al., 2016). For once our timing was perfect.

### From movement to Institute

In the following years, Eureka unfolded with an almost logarithmic speed. However, at the start, Eureka was met with some reluctance and even skepticism. Arguments against it varied from "we already have courses on translational medicine" to "why invest so much in such a small number of participants", and "we see no real problem here". But Salvo and I were used to far worse criticism, so this did not slow us down. Besides, we were not alone anymore: We had clear support from the other pioneers with whom we started Eureka in that basement in Siracusa back in 2008. Gradually it became clear that the success of the first course was not a lucky shot. The long-term impact on the participants became more evident (Weggemans et al., 2018). Consequently, we progressively gained more support culminating in institutional support from strong academic medical partners, starting with UMC Utrecht and Duke NUS. Next, other partners stepped in to support Eureka's mission to improve translational medicine: patient organizations, government, granting agencies, and more university medical centers. Thus, steadily, Eureka changed from a bottom-up initiative from a group of motivated individual scientists into a vibrant international network of universities, translational scientists, and patients, and closely linked to society. Remarkably we kept hearing the phrase "life and career changing" in many different settings and courses. So often that we started to perform educational research to better understand this apparent impact. Margot Weggemans, one of the PhD students studying translational scientists and the impact of Eureka, showed in a follow up study that the certificate course had a profound and lasting effect on the participants. She found that more than 85% of participants reported that Eureka changed the way they performed research. Most remarkably, this change persisted over time (Weggemans et al., 2018).

## Epilogue

This journey had an unexpected impact for me personally. It opened a whole new field of (educational) research— a type of research that turned out to be just as rich and complex as the biomedical research I was trained in. Janet Hafler, Olle ten Cate and Marieke van der Schaaf introduced me to the deeper layers of educational research, and I was mesmerized by what they told me. I started to read more and more about it, followed a post-master's course, and devoured books and articles on pedagogy psychology and the philosophy of education. Then, one day in 2017, to my own surprise I heard myself saying wholeheartedly yes when I was asked to apply for the position of vice dean for education at my home institution.

### Failure

The first basic idea for setting up Eureka came from the simple idea that others would need to learn from our failures. Though the program and learning objectives of Eureka ultimately became much more elaborate and more extensive, the idea of sharing of and learning from

failures stays deep in the DNA of Eureka. No quote better signifies the spirit of Eureka than the famous quote of Samuel Beckett: "Ever tried, ever failed? No matter. Try again. Fail again. Fail better" (Samuel Beckett quoted in Marshall, 2017). This may seem remarkable because in the highly competitive environment of life sciences, making failures can easily translate into being a failure. However, it is indeed crucial to admit to failures, to be unafraid to make them, and to learn from them and try again.

### Lessons Learned

It was impossible to foresee the chain of events we set in motion that morning in Paris in 2005. With hindsight, it is always easier to discern patterns that you did not see in the midst of the storm—or that may not have even been there. What we did aligned with the spirit of the time and resonated with then-unseen trends that eventually led to Open Science.

A few factors helped us along the way:

1. Conviction. We were so convinced — and so blinded by our idea — that we were simply too stubborn to give up.
2. Link to Practice. We were trained as clinicians and translational scientists, and we had experienced the Valley of Death ourselves.
3. Team. We sought out others to join us, and since that first meeting in 2008, we were no longer alone.
4. Fun. We genuinely enjoyed the journey.
5. Luck. No further explanation needed.

In closing, I realize now that we had started climbing what David Brooks calls the "second mountain" (Brooks, 2019). It was no longer about us: It was about creating impact for others. Looking back, this is what gave us the resilience to keep going.

P.S. I asked Salvo to proofread this article. He agreed with everything I wrote above but noted that I missed one important message: "The best is yet to come."

## References

Albani, S., Keystone, E. C., Nelson, J. L., Ollier, W. E., La Cava, A., Montemayor, A. C., Weber, D. A., Montecucco, C., Martini, A., & Carson, D. A. (1995). Positive selection in autoimmunity: Abnormal immune responses to a bacterial dnaJ antigenic determinant in patients with early rheumatoid arthritis. *Nature Medicine*, *1*(5), 448-452. https://doi.org/10.1038/nm0595-448

Albani, S., Koffeman, E. C., & Prakken, B. (2011). Induction of immune tolerance in the treatment of rheumatoid arthritis. *Nature Reviews Rheumatology*, *7*(5), 272-281. https://doi.org/10.1038/nrrheum.2011.36

Albani, S., & Prakken, B. (2006). T cell epitope-specific immune therapy for rheumatic diseases. *Arthritis & Rheumatism*, *54*(1), 19-25. https://doi.org/10.1002/art.21520

Benedictus, R., Miedema, F., & Ferguson, M. W. (2016). Fewer numbers, better science. *Nature*, *538*(7626), 453-455. https://doi.org/10.1038/538453a

Brooks, D. (2019). *The second mountain: The quest for a moral life*. Random House Publishing Group.

Calasan, M. B., van den Bosch, O. F., Creemers, M. C., Custers, M., Heurkens, A. H., van Woerkom, J. M., & Wulffraat, N. M. (2013). Prevalence of methotrexate intolerance in rheumatoid arthritis and psoriatic arthritis. *Arthritis Research & Therapy*, *15*(6), Article R217. https://doi.org/10.1186/ar4413

Calasan, M. B., & Wulffraat, N. M. (2014). Methotrexate in juvenile idiopathic arthritis: Towards tailor-made treatment. *Expert Review of Clinical Immunology*, *10*(7), 843-854. https://doi.org/10.1586/1744666X.2014.916617

de Jager, W., Prakken, B. J., Bijlsma, J. W., Kuis, W., & Rijkers, G. T. (2005). Improved multiplex immunoassay performance in human plasma and synovial fluid following removal of interfering heterophilic antibodies. *Journal of Immunology Methods*, *300*(1-2), 124-135. https://doi.org/10.1016/j.jim.2005.03.009

Delemarre, E., Roord, S., Wulffraat, N., van Wijk, F., & Prakken, B. (2011). Restoration of the immune balance by autologous bone marrow transplantation in juvenile idiopathic arthritis. *Current Stem Cell Research & Therapy*, *6*(1), 3-9. https://doi.org/10.2174/157488811794480726

Delemarre, E. M., Roord, S. T., van den Broek, T., Zonneveld-Huijssoon, E., de Jager, W., Rozemuller, H., Martens, A. C., Broere, F., Wulffraat, N. M., Glant, T. T., Prakken, B. J., & van Wijk, F. (2014). Brief report: Autolo-

gous stem cell transplantation restores immune tolerance in experimental arthritis by renewal and modulation of the Teff cell compartment. *Arthritis & Rheumatology*, *66*(2), 350-356. https://doi.org/10.1002/art.38261

Kamphuis, S., Hrafnkelsdottir, K., Klein, M. R., de Jager, W., Haverkamp, M. H., van Bilsen, J. H., Albani, S., Kuis, W., Wauben, M. H., & Prakken, B. J. (2006). Novel self-epitopes derived from aggrecan, fibrillin, and matrix metalloproteinase-3 drive distinct autoreactive T-cell responses in juvenile idiopathic arthritis and in health. *Arthritis Research & Therapy*, *8*, Article R178. https://doi.org/10.1186/ar2088

Kamphuis, S., Kuis, W., de Jager, W., Teklenburg, G., Massa, M., Gordon, G., Boerhof, M., Rijkers, G. T., Uiterwaal, C. S., Otten, H. G., Sette, A., Albani, S., & Prakken, B. J. (2005). Tolerogenic immune responses to novel T-cell epitopes from heat-shock protein 60 in juvenile idiopathic arthritis. *The Lancet*, *366*(9479), 50-56. https:/ / doi.org/ 10.1016/ S0140-6736(05)66827-4

Koffeman, E. C., Genovese, M., Amox, D., Keogh, E., Santana, E., Matteson, E. L., Kavanaugh, A., Molitor, J. A., Schiff, M. H., Posever, J. O., Bathon, J. M., Kivitz, A. J., Samodal, R., Belardi, F., Dennehey, C., van den Broek, T., van Wijk, F., Zhang, X., Zieseniss, P., … Albani, S. (2009). Epitope-specific immunotherapy of rheumatoid arthritis: Clinical responsiveness occurs with immune deviation and relies on the expression of a cluster of molecules associated with T cell tolerance in a double-blind, placebo-controlled, pilot phase II trial. *Arthritis & Rheumatism*, *60*(11), 3207-3216. https://doi.org/10.1002/art.24916

Lewis, G. F., & Brubaker, P. L. (2021). The discovery of insulin revisited: Lessons for the modern era. *The Journal of Clinical Investigation*, *131*(1), Article e142239. https://doi.org/10.1172/JCI142239

Nescio (2012). *Amsterdam Stories*. (D. Searls, Trans.) New York Review Books. (Original work published 1933)

Marshall, C. (2017, December). Samuel Beckett's mantra: Try again, fail again, fail better. *Goethe Institute USA, Los Angeles.* https://www.goethe.de/ins/us/en/sta/los/bib/feh/21891928.html

Prakken, B. J., Roord, S., van Kooten, P. J., Wagenaar, J. P., van Eden, W., Albani, S., & Wauben, M. H. (2002). Inhibition of adjuvant-induced arthritis by interleukin-10-driven regulatory cells induced via nasal administration of a peptide analog of an arthritis-related heat-shock protein 60 T cell epitope. *Arthritis & Rheumatism*, *46*(7), 1937-1946. https:/ / doi.org/10.1002/art.10366

Prakken, B. J., Samodal, R., Le, T. D., Giannoni, F., Yung, G. P., Scavulli, J., Amox, D., Roord, S., de Kleer, I., Bonnin, D., Lanza, P., Berry, C., Massa, M., Billetta, R., & Albani, S. (2004). Epitope-specific immunotherapy induces immune deviation of proinflammatory T cells in rheumatoid arthritis. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(12), 4228-4233. https://doi.org/10.1073/pnas.0400061101

Trentham, D. E., Dynesius-Trentham, R. A., Orav, E. J., Combitchi, D., Lorenzo, C., Sewell, K. L., Hafler, D. A., & Weiner, H. L. (1993). Effects of oral administration of type II collagen on rheumatoid arthritis. *Science*, *261*(5129), 1727-1730. https://doi.org/10.1126/science.8378772

van Eden, W., Thole, J. E., van der Zee, R., Noordzij, A., van Embden, J. D., Hensen, E. J., & Cohen, I. R. (1988). Cloning of the mycobacterial epitope recognized by T lymphocytes in adjuvant arthritis. *Nature*, *331*(6152), 171-173. https://doi.org/10.1038/331171a0

Weggemans, M. M., van der Schaaf, M., Kluijtmans, M., Hafler, J. P., Rosenblum, N. D., & Prakken, B. J. (2018). Preventing translational scientists from extinction: The long-term impact of a personalized training program in translational medicine on the careers of translational scientists. *Frontiers in Medicine*, *5*, Article 298. https://doi.org/10.3389/fmed.2018.00298

Weiner, H. L., Friedman, A., Miller, A., Khoury, S. J., al-Sabbagh, A., Santos, L., Sayegh, M., Nussenblatt, R. B., Trentham, D. E., & Hafler, D. A. (1994). Oral tolerance: Immunologic mechanisms and treatment of animal and human organ-specific autoimmune diseases by oral administration of autoantigens. *Annual Review of Immunology*, *12*, 809-837. https://doi.org/10.1146/annurev.iy.12.040194.004113

Zonneveld-Huijssoon, E., van Wijk, F., Roord, S., Delemarre, E., Meerding, J., de Jager, W., Klein, M., Raz, E., Albani, S., Kuis, W., Boes, M., & Prakken, B. J. (2012). TLR9 agonist CpG enhances protective nasal HSP60 peptide vaccine efficacy in experimental autoimmune arthritis.

*Annals of the Rheumatic Diseases*, *71*(10), 1706-1715. https://doi.org/10.1136/annrheumdis-2011-201131

**Correction notice**
Incorrect Special Issue Labeling (Article erroneously excluded): This article was previously not labeled as part of a special issue due to an error. This has now been corrected.

# Issues in Clinical Studies Leading to Medical Research Ethics Committee (MREC) Negative Decisions

Sigrid E. M. Heinsbroek[1], Vincent Bontrop[1], Rutger P. Chorus[1], C. Michel Zwaan[1]

The rationale behind a Medical Research Ethics Committee (MREC) negative decision is always shared directly with the applicants. However, insight into the review process and common reasons for a negative decision may also be valuable for other researchers, clinical research organizations and people with an interest in MREC processes. To our knowledge Medical Research Ethics Committees (MRECs) do generally not report on the negative decisions they issue and on the underlying rationale for such decisions. Here we give insight into the MREC review process by briefly describing procedures and discussing the negative decisions issued by MREC NedMec in the past five years.

**Keywords** *clinical trials, clinical studies, research assessment, medical research ethics committee, medical research ethics*

[1] METC NedMec, Utrecht, the Netherlands

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
October 12, 2023
**Accepted**
January 11, 2024
**Published**
March 22, 2024

**Correspondence**
METC NedMec
s.e.m.heinsbroek@umcutrecht.nl



In the Netherlands, medical scientific research in which human volunteers are subjected to procedures, or are required to follow rules of behavior, needs to be reviewed by an accredited MREC, as per the Dutch law for 'Medical Research involving Human Subjects'. The Netherlands has 14 accredited MRECs and their operations are overseen by the Central Committee on Research Involving Human Subjects (CCMO).

MRECs determine if studies comply to the current legislation, regulations and guidelines concerning human subject research. This includes the General Data Protection Regulation (GDPR; European Parliament and the Council of the European Union, 2016), the Medical Research Involving Human Subjects Act (WMO; Nederlandse Overheid, 2022), the Clinical Trial Regulation (CTR; European Parliament and the Council of the European Union, 2014) or the Clinical Trial Directive (CTD, till 31st of January 2022; European Parliament and the Council of the European Union, 2001), Medical Device Regulation (MDR; European Parliament and the Council of the European Union, 2021), In Vitro Diagnostics Regulation n (IVDR; European Parliament and the Council of the European Union, 2017), guideline for Good Clinical Practice (European Medicines Agency, 2016), Dutch Embryo Act (Nederlandse Overheid, 2021) and the Declaration of Helsinki (World Medical Association, 2013). The kind of documents needed for assessment, the way a dossier is submitted, and the committee that assesses a research protocol depends on legislation and the regulations applicable to the study. With the increase in regulation, it becomes more common that submitted studies are returned to the applicant because of mistakes in the submission procedures or documents. For instance, clinical trials with medicinal products are sometimes still submitted directly to the MREC, while these should now be submitted via the Clinical Trials Information System (CTIS). Furthermore, medical device studies are not always submitted as an MDR-study, and as a result documents may be missing. Only when the

study is submitted correctly and all essential documentation is available, the review process can start.

## MREC review process

In the Dutch review process, medical ethical, regulatory and scientific assessment is integrated leading to a thorough review system which is, to the best of our knowledge, unique in the world. A submitted study will be reviewed in a plenary MREC committee meeting where the presence of the following experts is mandatory:

- a medical doctor,
- a lawyer,
- an ethicist,
- a methodologist/statistician,
- a patient advocate specifically for the perspective of the human subject.

For specific cases, further experts are required to be part of the committee:

- a pediatrician (obligatory for research conducted with minors),
- a clinical pharmacologist (for clinical trials with medicinal products),
- a hospital pharmacist (for clinical trials with medicinal products),
- a medical device expert (for clinical investigations with medical devices).

Every expert meets strict requirements and is approved by the CCMO before taking part in an MREC (CCMO, 2020). Together the committee assesses the study by weighing benefit versus participant burden and risk. They take into account:

- if it is likely that the scientific research will lead to the discovery of new insights in the field of medical science,
- if there are alternatives to the submitted study available that reduce burden or risk of participants, and
- if the interests of the subject or other current or future patients are served by the research.

The committee pays close attention to safety and ethical aspects. The latter includes the informed consent procedure in which it is of upmost importance that participants are clearly informed of the purpose and the potential benefit but also the risks of the study. Contracts will be reviewed based on CCMO guidelines on the review of research contracts (CCMO, 2011) and should not contain unreasonable restrictions regarding the disclosure of the results, or premature termination of the study, or unreasonable financial incentives. Furthermore, it should be clear that the study design enables researchers to answer the research question, that enough study participants can be recruited and included, and that the available funds will cover the costs of the whole study to completion.

The review process of MREC NedMec is depicted in figure 1. After the first review by the committee, questions are sent to the investigators. Depending on the kind of questions asked, the further assessments will either be done in a plenary meeting when the committee requires substantial modifications or will be delegated to an executive committee when only minor issues need to be resolved (fig.1).

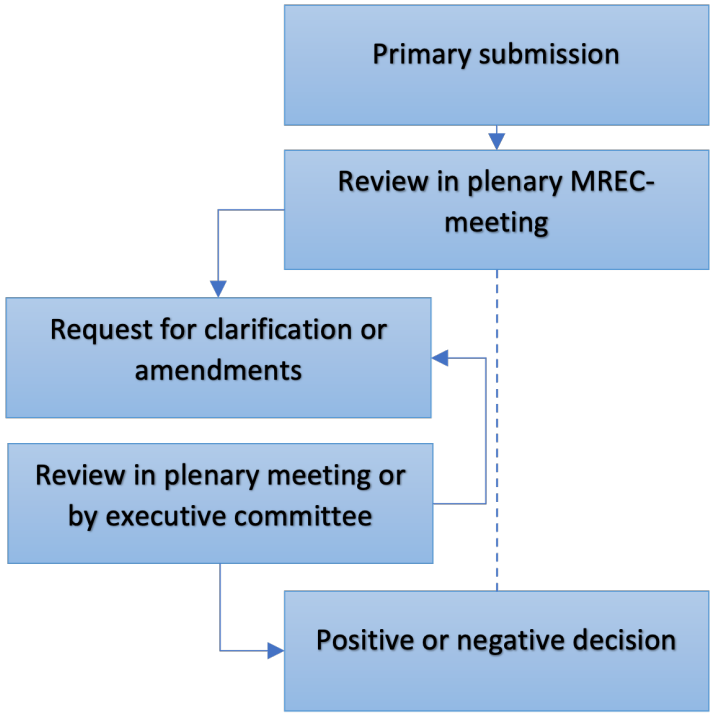



**Figure 1** *Stages of the MREC review process.*

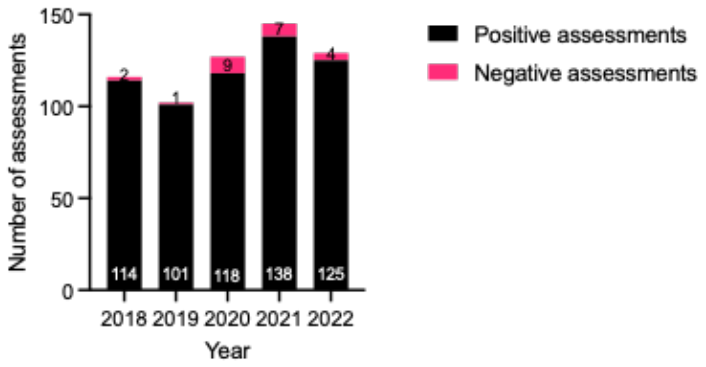



**Figure 2** *Number of primary assessments by METC NedMec over five years (2018 to 2022).*
The number of positive and negative decisions made by METC NedMec over the course of five years.

## Negative decisions

Negative decisions are relatively uncommon. In the last three years, only 1.7% of the studies that have been assessed in the Netherlands by an accredited MREC or the CCMO received a negative decision (CCMO, 2023) . A negative decision is rarely given after the first MREC review of the study. Depending on whether the applicable legislation offers the possibility to request additional information and modifications multiple times, generally two or more rounds of discussion take place before a negative decision is issued. Under CTR this is limited to one round of questions.

Over the last five years MREC NedMec (known as METC Utrecht until 2021) assessed 596 studies, of which 23 received a negative decision (fig. 2). The number of negative decisions peaked in 2020, which was most likely due to a large number of COVID-19 studies that were submitted under time pressure (METC Utrecht, 2020).

In the next paragraph we will discuss the reasons that led to negative decisions by MREC NedMec in the last five years (2018 to 2022) regarding studies that are subject to the WMO. These are summarized in Figure 3.

### Incomplete research file

In this period most negative decisions were given due to an incomplete research file despite rounds of questions. For instance, when a protocol is submitted as a WMO-study while a medicinal product or a medical device is used, one may forget to submit quality or safety information on the medicinal product or device.

### Scientific validity or safety

It occasionally happened that scientific justification remained incomplete, and in most of these studies the lack of sufficient quality or safety information led to a negative decision. For example, it happened that a study was submitted as a phase IB/IIA study, while the requirements for a phase I and II study are different, and information on the data safety monitoring board, safety criteria and post-trial access were missing, and not handled well in the answer to the committee review. When safety information was sufficient, the intervention could be considered too burdensome compared to the scientific contribution, which led to negative decisions in three studies.

### Benefit/risk assessment

Surprisingly, in more than 20% of the negative decisions, researchers could not sufficiently substantiate the burden and risks for the study participant. In these studies, the MREC's questions about necessity of the study were often answered briefly without a balanced benefit/risk assessment. For example, under Dutch law it is forbidden to conduct scientific research with children under 16 years of age unless these children perceive some benefit from the study, or when the study cannot be carried out without their cooperation (Nederlandse Overheid, 2022). Two studies received a negative decision because the added scientific value for including children remained unclear. For one study the committee was convinced of its relevance, but insurance for the children to cover any potential damage incurred as a result of participating in the study could not be arranged. This also resulted in a negative decision.

### Consent procedure and study design

Despite rules and guidelines on how informed consent should be obtained (CCMO), two studies received a negative decision due to an un-

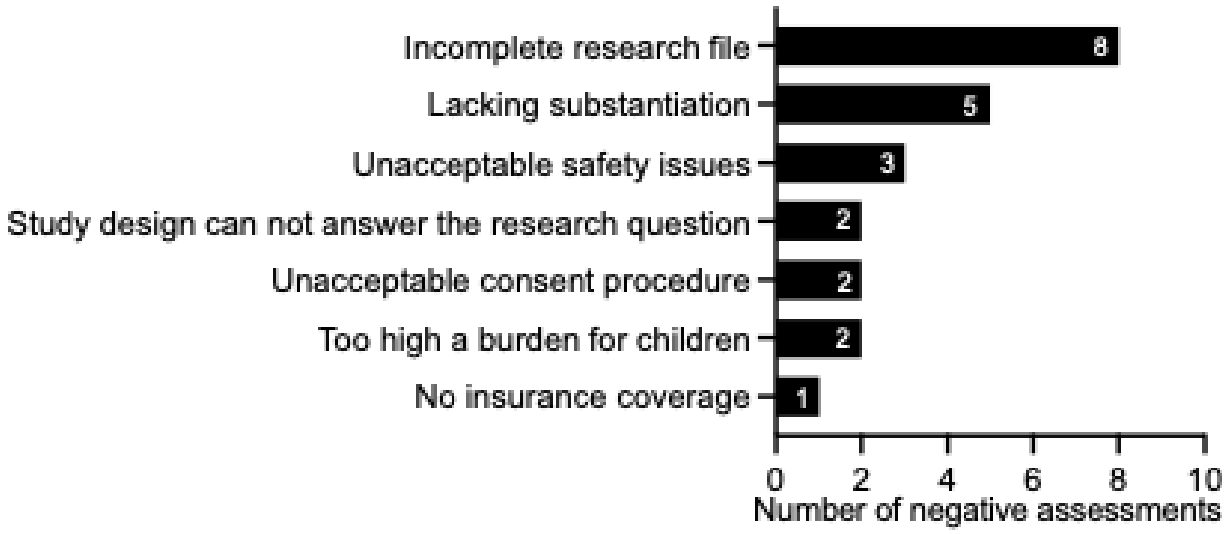



**Figure 3** *Reasons for negative decisions by MREC NedMec over five years (2018 to 2022).*

acceptable procedure to obtain informed consent. In one study researchers deemed it unnecessary to ask for consent, which the committee did not agree with. In the other study researchers intended to directly phone potential study participants for research purposes, without knowing if they were willing to participate. Some studies received a negative decision based on improper study design. For example, results could not be directly attributed to the intervention because of a missing control group, or a subjective primary endpoint was used which was more likely to give an unreliable answer to the study question.

### Appeal

After receiving a negative decision most studies were modified and resubmitted for a new review, but there is also the possibility to appeal the MREC decision at the CCMO. Less than half of the negative decisions were appealed (fig 4). Of the six studies that were appealed, two had received a negative decision due to the consent procedure, one due to the burden for children, one due to safety issues, and one due to missing substantiation of the risk-benefit assessment. Three appeals were withdrawn after discussion with the CCMO. Three other appeals received an unfounded verdict by the CCMO after appeal, meaning that the MREC decision remained valid and unchanged. In the one study where the appeal was founded, researchers submitted new information which affected the risk-benefit assessment and helped in the final decision. This might also have led to a positive decision by METC NedMec, if the information was available at the time of review. Together this shows that while appeal is an option, investigators may be better off by adjusting the study or providing the information asked for by the MREC.

## Recent changes in regulation and their consequences

Until the introduction of the ECTR, the role of the competent authorities in the Netherlands was limited, and the assessment was done in a decentralized fashion by the accredited MRECs. These would assess the entire dossier (now referred to as part I and part II under the ECTR). Currently, studies subject to the CTR or certain articles of the MDR and IVDR will have a validation period in which completion of the study file is checked and missing information is requested. This should prevent negative decisions caused by submitting an incomplete

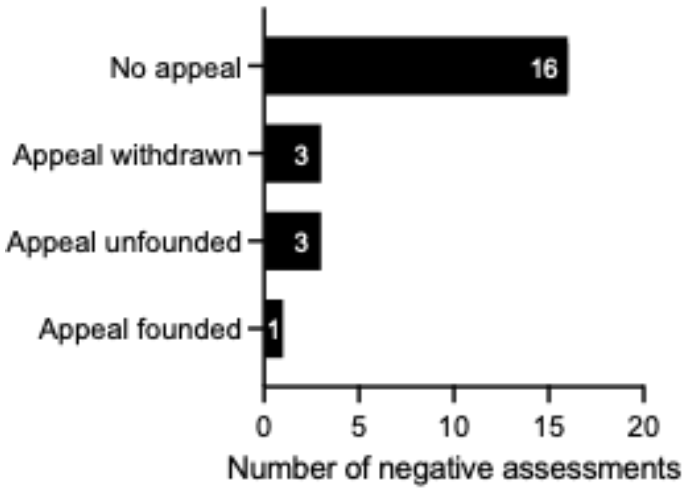



**Figure 4** *Appeal at the Central Committee on Research Involving Human Subjects (CCMO)*
Number of studies that received a negative decision by METC NedMec between 2018 and 2022. The majority of decisions were not appealed. When the decision was appealed there were three possible outcomes: 1) appeal founded, 2) appeal unfounded, 3) appeal was withdrawn after a first conversation with the CCMO.

dossier. However, since the CTR came into effect on January 31, 2022, it seems that the number of negative decisions for clinical trials with medicinal products may be increasing. In the last year, METC NedMec issued three negative decisions out of 31 studies assessed under CTR (9,6%). In the year before, two out 55 clinical trials with medicinal products assessed under CTD received at negative decision (3,6%). In our opinion, the following factors contribute to a higher number of negative decisions:

- Researchers are still unfamiliar with the CTR and the different available guidelines, resulting in research files that do not meet the requirements that apply under the CTR.
- Deadlines are tighter, and the 12 days to comply is often too short for investigators to make major changes when information or whole documents are missing.
- Studies subjected to the CTR only have one round of review, after which a decision needs to be made.
- Particularly in assessments of international studies, where one country is appointed as the reporting member state, it happens that questions deemed essential by a concerned member state are removed in the communication to investigators. This leads to negative decisions on conducting the study in The Netherlands.

Together these changes will most likely increase the number of MREC negative decisions further in the future. Furthermore, appeal against a negative decision under CTR at the CCMO is not always possible; in that case the study needs to be resubmitted.

## Conclusion

Medical ethical and scientific review is strongly integrated in the Dutch MREC process leading to a thorough assessment. Negative decisions are uncommon, and researchers can support quick and positive MREC review by submitting complete and well-substantiated studies in the correct manner, according to the regulations applicable to the study. It also helps when questions from the committee are answered in a clear and thorough manner and the necessary adjustments are made. When uncertain while designing a study, we recommend contacting an ethicist or methodologist or asking for scientific or regulatory advice. Scientific guidelines for studies with medical products can be found on the European Medicines agency website and information for clinical studies with medical devices can be found on the webpage of medical device coordination group. Guidelines on documents needed for a complete research file are available at the CCMO's website and when in doubt, one can always ask the MREC secretariats for advice.

## Abbreviations

| | |
|---|---|
| CCMO | Central Committee on Research Involving Human Subjects |
| CTR | Clinical Trial Regulation |
| CTIS | Clinical Trials Information System |
| GDPR | General Data Protection Regulation |
| IVDR | In Vitro Diagnostics Regulation |
| MDR | Medical Device Regulation |
| MREC | Medical Research Ethics Committee |
| WMO | Medical Research Involving Human Subjects Act |

## Acknowledgments

We thank Myriam van der Loo, Jan Paul de Boer and Bianca Goemans for commenting on the draft of this paper.

## References

CCMO. (2011). Herziene CCMO-richtlijn beoordeling onderzoekscontracten. https://www.ccmo.nl/ over-de-ccmo/ publicaties/ richtlijnen/ 2011/ 09/ 20/ ccmo-richtlijn-beoordeling-onderzoekscontracten

CCMO. (2020). Richtlijn van de centrale commissie mensgebonden onderzoek, de CCMO. https:/ / www.ccmo.nl/ metcs/ publicaties / richtlijnen/ 2019/ 11/ 14/ ccmo-richtlijn-deskundigheidseisen-metc-leden

CCMO. (2023). CCMO jaarverslag 2022 de CCMO in beweging - in Nederland en Europa. CCMO. https://www.ccmo.nl/over-de-ccmo/publicaties/jaarverslagen/2023/03/13/jaarverslag-ccmo-2022 European Medicines Agency. (2016). Guideline for good clinical practice. https://www.ema.europa.eu/en/ich-e6-r2-good-clinical-practice-scientific-guideline

European Parliament and the Council of the European Union. (2001). Directive 2001/20/ec of the European Parliament and of the Council of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use. https:/ / eur-lex.europa.eu/ legal-content/EN/TXT/?uri=CELEX%3A02001L0020-20220101

European Parliament and the Council of the European Union. (2014). Regulation (EU) no 536/2014 of the European Parliament and of the Council of 16 april 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/ec. https:/ / eur-lex.europa.eu/ legal-content/ EN/ TXT/ ?uri=CELEX:02014R0536-20221205

European Parliament and the Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) Official Journal of the European Union. https://gdpr-info.eu/

European Parliament and the Council of the European Union. (2017). Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro di-

agnostic medical devices and repealing Directive 98/79/ec and Commission Decision 2010/227/EU. https://eur-lex.europa.eu/legal-content/ EN/ TXT/ ?uri=CELEX%3A32017R0746&qid=1687428283895

European Parliament and the Council of the European Union. (2021). Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) no 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. https:/ / eur-lex.europa.eu/eli/reg/2017/745/2020-04-24

METC Utrecht. (2020). Jaarverslag 2020 METCUtrecht.

Nederlandse Overheid. (2021). Embryowet. https://wetten.overheid.nl/BWBR0013797/2021-07-01

Nederlandse Overheid. (2022). Wet medisch-wetenschappelijk onderzoek met mensen. https://wetten.overheid.nl/BWBR0009408/2022-07-01

World Medical Association. (2013). WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects. https:/ / www.wma.net/ policies-post/ wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

# The "Function" of Art?: Challenges of Setting Up Artistic Research Residencies in Elderly Care Institutions

Falk Hübner[1], Gjilke Keuning[2], Marijke Lucas[1]

In this article, the authors reflect on the co-creative processes of three artistic research residencies in an elderly care institution in Leiden, The Netherlands. The artists were asked to immerse themselves in the institution for half a year, 2 days a week, to carry out practice-based and experience-driven research. Through artistic methods such as sketching, photographing, or creating mockups, the artists were supposed to reflect on their experiences in the institution, in particular on the relation of the residents to staff, caretakers, and their loved ones at home. These reflections were supposed to lead to final artistic works that offer alternative views on the institution and everything happening in it, from the inside out. Two challenges became apparent in these residencies: First, balancing the imaginative and speculative artistic process with the urge to "help", to make a meaningful — however instrumental — contribution to the institution on the other. Second, the urge to use spoken language (i.e. dialogue and discussion) quickly dominated the exchange between artists and participants in the institution and threatened the work with artistic materials, which are often not lingual in nature. The institution itself and its staff are highly thankful for the fresh and sometimes unexpected/unorthodox views of the artists, but do these more 'systemic' observations actually help the artistic process? And what exactly is the value that the arts and artists have to offer with their practices?

**Keywords** *artistic research, art residencies, arts and healthcare, co-creation*

[1] Fontys Academy of the Arts, Tilburg, the Netherlands
[2] HKU University of the Arts, Utrecht, the Netherlands

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
October 23, 2024
**Accepted**
April 2, 2025
**Published**
June 4, 2025

**Correspondence**
Fontys Academy of the Arts
f.hubner@fontys.nl



## Introduction

In this article, we reflect on the co-creative processes of three artistic research residencies in Topaz, an elderly care institution in Leiden, The Netherlands. This project, called "The Art of Bringing Together" (Dutch original: "De kunst van het samen brengen"), was a collaboration between elderly care institution Topaz in Leiden, HKU University of the Arts Utrecht, and Fontys Academy of the Arts Tilburg. Three artists were asked to immerse themselves in this context for half a year, 2 days a week, in order to carry out practice-based and experience-driven research. Through artistic methods such as sketching, photographing, or creating mockups [1], the artists were supposed to reflect on their experiences in the institution, in particular on the relation of the residents to staff, caretakers, and their loved ones at home. These reflections should lead to final artistic works that offer alternative, imaginative or speculative views on the institution and what happens in it, from the inside out.

### Alienation in the process of moving from home to care home

The point of departure of the research project was the question *how it is possible to create space for that which makes life worth living*. This question quickly proved too open and

[1] For a more substantial catalogue of artistic methods see Badura et al. (2015). For a conceptual framework of method, research strategy and methodology in artistic research, see Hübner (2024, pp. 31-39).

## Take-home message

We concluded that the residency at the Topaz retirement home was not appropriate for research into residents' relationships with loved ones at home. When artists are offered an open way to connect with the social context, they can find ways that resonate with the underlying needs of older people and loved ones. Artistic interventions and expressions respond to questions in social contexts in powerful, imaginative, speculative, and unexpected ways. We argue that being "functionless" offers an essential value: It allows art and artists in healthcare to actually flourish.

too large, needing more specificity. Therefore we refocused the project on the challenges Topaz was grappling with, in particular the period when the residents move from home to care home. Topaz wondered:

> *How can people who grow older stay focused on what is meaningful for them in life — together with their loved ones, friends and family — despite chronic health-issues?*[2]

Topaz is a care home for elderly people who cannot take care of themselves anymore, in the broadest sense and for a variety of reasons. Staff members at Topaz have observed and experienced how difficult the transition from home to care home is, for the residents themselves as well as for partners, friends, and family. Their homes, most of their belongings, surroundings, routines, and relationships get lost, and relationships that do not get lost tend to change drastically.

Rather than seeing that existing relationships continue, Topaz observed a process of alienation.[3] The institution necessarily takes over most of the care, which can feel ambiguous for close relatives who have offered this care up to the point of moving. They can feel as though they are falling short, because they are no longer able to provide the necessary care. Now that their loved one has moved to a care facility, they find themselves in a very different position and are unsure of what they can and should do. They don't want to be in the way, but they are still concerned and worried about their loved one and still want to spend time with them. In this, they feel dependent on the nurses and care professionals. Due to this sense of dependency, they do not dare to ask everything or be critical.

The professionals, on the other hand, mainly focus on their relationship with the new resident and often do not know how (or have insufficient time) to deal with the close relatives. They do not want to burden the resident's loved ones, who are often exhausted. They also feel responsible for the quality of care, which they are trained and hired for. It is not self-evident to leave caring tasks to family members, even though they have often done this for a long time. Additionally, doing it themselves is often faster, easier, and thus more efficient than continually coordinating with loved ones about who does what. Small initiatives to facilitate more connection with friends and family only develop slowly, due to existing behavior patterns and time pressure, among other reasons.

For the people moving to a care facility (the residents), it can be a great loss when relationships vanish or change. This is due to two primary reasons: First, relationships are valuable in themselves; they matter, especially during a significant event such as moving. Second, relationships make various important things possible, such as doing enjoyable activities together. There are things that professionals simply cannot take over, but which are lost in the almost automatic process of "taking over everything." And, finally, for the relational activities that professionals and volunteers can take over, there is an increasing scarcity in the available time to perform these tasks.

### "Calling for the arts" — Research question, strategy, and the approach of artistic research residencies

By collaborating with the arts, Topaz hoped to create room to find new experiences and al-

[2] The original question in Dutch was: Hoe kunnen mensen bij het ouder worden zich blijven richten op wat er voor hen toe doet - samen met de mensen die voor hen belangrijk zijn - ondanks chronische gezondheidsuitdagingen?

[3] Such observations and analyses were shared with the project team and the artists during preparatory conversations and a kick-off meeting (see 'The overall process — Getting started [...]').

ternatives, which might provide alternatives to existing patterns. The initial research question was posed mainly by the care institution and had the following form:

> *How can we contribute to keeping meaningful relations in and after the period of moving (from home to care institution)?*[4]

This question was based on the assumption that such relations (with families, partners, or friends) are valuable in themselves and deserve to be kept intact. How can such relations be continued in a process of reduced self-reliance, either through age or health issues? We chose to approach these questions in and through artistic residencies[5] and, within them, processes and activities of artistic co-creation. Three artists spent half a year, 2 days per week, in residence in one of the departments at Topaz, using the period of moving "from home to care home" as a lens to work with. By being present and in contact with several actors (people living there, working there, voluntary staff, families and friends), the artists experienced what happened to the relations of the people who have moved. They used their experiences as inspiration to react by means of (co-)creating artistic work. A co-creative approach means that the activities the artists carry out includes the people this work is about; the artists collaborate with them as co-creators. This also means that the artistic work ideally is not created *for* or *about* the context and people in this context, but *with* them. Such artistic processes are characterized by multisensorial, imaginative, and speculative ways of working, and typically lead to new space to develop emerging ideas, feelings and perspectives. The artists let others participate in their experience. In doing so, this work might be able to create a different, more connective perspective on life and care.While the main actors in this approach were the artists (young graduates from two Dutch art academies: HKU University of the Arts Utrecht and Fontys Academy of the Arts Tilburg[6]), there were a number of people involved in facilitating, supervising, and working with them: Project leader "Kleinschalig Zorgen" Eva van Zelm coordinated the project at Topaz and was the first point of contact for the artists during their residence. The artists were further supervised by a team from all three participating institutions: trailblazer Arts, Health, and Society Gjilke Keuning and media artist and pedagogue Erwin Slegers (both HKU),[7] as well as artist-researcher and professor of Artistic Connective Practices Falk Hübner (Fontys). At Topaz, residents, staff, team leaders, and volunteers were also involved. In the final stage of the residencies, industrial designer Marijke Lucas joined the team as a researcher to provide another perspective to the project. Marijke interviewed the artists extensively after the project was finished (R. Krijgsman, personal communication, November 28, 2023; R. Tavoraite, personal communication, November 29, 2023; M. Pietjouw, January 23, 2024). All three artists have read a draft of this article and provided full consent to use their names and professional identities.

In summary, in this project we were mainly interested in developing new perspectives on the life of the elderly with their loved ones, while and after moving from their home into a care institution. At the same time, we aimed to collect and explore new perspectives on working in and through the arts in the area of healthcare — through the methodological approach of artistic residencies.

## The overall process

### Getting started and arriving at the organization

The preparation of the project included composing the core team and supervision structure, making contracts, and formalizing

[4] Dutch original: Hoe kunnen we zorgen dat waardevolle relaties in en na de verhuisperiode behouden blijven?

[5] The approach of a residency is closely related to (auto-)ethnographic methodologies and action research: The artists spend a substantial part of their time in the care institution, with the residents and staff, and a large part of the creation process takes place there. The artists are thereby enabled to experience this context from within, and thus 'to gain insight by being in the same social space as the subjects of their research' (Madden, 2017, p. 1).

[6] HKU en Fontys have curated the participating artists together, in consultation with Topaz. Together the entire team was responsible for matching the artists to the teams at Topaz and their departments.

[7] Slegers has also worked in geriatric contexts, has experience as caregiver, and is regularly involved in extracurricular projects.

informed consent regarding privacy and data storage. Healthcare staff, residents, and families were informed about the project and were asked for consent to participate.

Three artists were invited to take part in this project: visual and performance artist Rimanta Tavoraite, designer Mandy Pietjouw, and scenographer and multidisciplinary artist Robin Krijgsman.

The artists met the organization at a kick-off meeting on location, where Topaz staff introduced them to the different locations and the diversity of residents, such as dementia patients and people with somatic complaints or issues of independence (or the lack thereof). The whole team (including Topaz staff, supervisors, and the artists) visited the locations to get a more concrete and practical impression. Based on this information and the visits, the artists chose three separate locations (in exchange with Topaz), different in size and complexity of care that residents need. Mandy went to work at a location where residents with more advanced forms of dementia live, whilst Robin worked at a location for temporary housing, and Rimanta went to a location with residents with (acquired) physical disabilities. To get to know, and get used to the context, residents, and staff, the artists participated in the work in the various departments: They helped with setting up tables, brought residents to their rooms, took walks with residents outside, or just talked with them. At the same time, this was the start of the creative process, which in total consisted of four stages:

1. Observing, experiencing, and participating,
2. Generate ideas on the basis of these lived experiences,
3. Reflection and proposition for a work (in whatever kind or form) that meaningfully contributes to the context,
4. Public manifestation of the work and sharing of the outcomes.

While these four stages had been designed for this particular project, they are based on both earlier artistic research work in relation to healthcare in our own institutions (see Dörr & Hübner, 2017), as well as loosely related to what Anke Coumans (2020) proposes as the first four of five settings of a "designing attitude" in research projects related to dementia care setting: *anecdotal, observing, situating and in scenario*. The five settings of a "designing attitude" build on Isabelle Stengers' concept of an Ecology of Practices, which aims for "new possibilities for practices to be present or to connect" (Coumans 2020, n.p., our translation).

The core team came together for collective reflection moments four times during these stages, next to weekly individual coaching moments, both online and on site. The core team and staff of the institution met three times: for the kick-off, a mid-process sharing and reflection session, and the public manifestation.

## In-between and final reflections

The first reflection meeting with the core team, heads of departments, and other interested staff took place after the first few months. In this meeting, the artists shared their experiences and ideas that had emerged up to this point. The session finished with a "Rich A4", a reflective work form developed by the HKU professorship Art and Professionalization. In this work form, participants answer a number of reflective, creative, and associative questions on a piece of paper, including drawing and thinking about a title for one's reflections, for example.

This specific session showed a clear tension and uneasiness. The artists shared their experiences and reflections mainly through telling, using reflective and discursive spoken language (judgmental at times), which resulted in defensive behavior on the side of staff. One of the artists, Rimanta, offered artistic material and a conversation through the metaphor of "Topaz as theater," leading to questions such as: "How does this theater work?" Or: "Which role has the audience in this theater?" This approach proved to be more open and playful and resulted in a much more lively discussion.

Furthermore, the coaches, artists, and coordinators had three moments of reflection on the project in the form of online sessions. Different forms of reflections were possible. For example, Mandy chose the form of a poem in one of the sessions, a poem she also read during the final presentation.

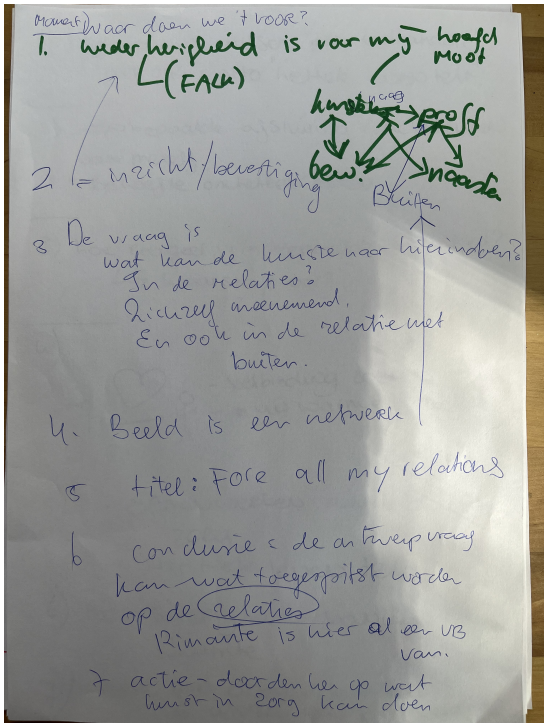



**Figure 1** A "rich A4", created by one of the participants (anonymized). Image taken by Gjilke Keuning.

**Figure 2** Mandy carries out last adjustments at the plant. Image taken by Gjilke Keuning.

## Final works

The final works were presented and shared on one day, organized by and at Topaz in the "living room", a huge space in which residents, families, friends, and staff can come together. Both staff and residents were invited, and the entire team was present, as were the artists. The artists presented their works, which led to several reactions and dialogues. The works were diverse concerning materials, aesthetics, form, or how the works related to the surroundings, people in the institution, how they interacted... All of these parameters were different.

Mandy's final presentation consisted of two works: A "feel-cloth/-fabric" as a first, smaller study, and a plant made of felt and an inner steel construction as larger work. The plant works in and of itself: Its central idea is to provoke conversations by "standing in the way" and thus bringing residents out of their ordinary rhythm. Residents passed the plant when walking into the living room and were impressed by the craftsmanship, by the soft feel of the plant, or simply by "how beautiful it was."

Conceived as a more general reflection on the healthcare context, Robin's work presented an object in and of itself, as well as a model of an imaginary space. The work reflects on the institution's process of moving into new buildings, and poses the question what happens when furniture is taken away from a place and put into a new place, into a new structure. Which opportunities exist to position things differently, or would one choose for old and familiar structures?

Rimanta's work, a hybrid between exhibition, performance, and workshop, worked activating for both residents and staff. She used photos of residents as well as her own childhood and combined two photos in one, through a technique of weaving (see Figure 5). The photos sparked memories amongst the older residents in particular and led to engaged conversations, with residents telling stories from their lives. In the interactive part, Rimanta

**Figure 3** Mandy's "feel-cloth". Image taken by Gjilke Keuning.

**Figure 4** Robin's model with furniture hanging from the ceiling. Image taken by Gjilke Keuning.

invited the residents and staff to work with various objects themselves, a kind of playful artistic crafting.

## Challenges, dilemmas, issues, topics — A catalogue

### General issues/challenges on co-creation

As in large parts of the healthcare sector, staff regularly changed, worked in different shifts, and on changing days. This presented a general challenge for the artists. There was always a different, new person working in a department and it was difficult to build personal connections. These constant changes made it hard for the artists to explain and clarify what they as artist actually do, how they think and work. It seemed difficult for staff to understand how an artistic process works (and how should they know?), which also resulted in uncertainty regarding mutual expectations.

The approach of co-creation proved to be a bigger challenge than we expected. In retrospect, none of the three artists has truly worked co-creatively in the sense of involving

**Figure 5** The materials Rimanta offered to work/craft with. See top left for the interwoven photos. Image taken by Gjilke Keuning.

**Figure 6** Two residents in conversation about a photo. Image taken by Gjilke Keuning.

**Figure 7** A resident working with Rimanta's materials. Image taken by Gjilke Keuning.

participants (residents or staff) as partners who more or less create artistic work: Rimanta came closest to this idea, as she built her work by visiting residents and collecting personal stories, experiences, as well as childhood images of residents and staff, and composing new images from them. Mandy, in contrast, "tested out" her ideas and iterations of the work, more comparable to a designer-approach, bringing her designs and iterations into exchange with staff and residents. Robin, instead, appeared to take the position of an outside observer, and took these experiences into his own studio to work. He rather autonomously created — as a comment on this context, so to speak. Concerning the side of the care institution, the staff and residents had no actual reference or frame for artistic practices, which made it a challenge to involve them. We think a guest studio would have been a possible solution for the artists to create their own surrounding within the institution, their own world in which they could invite staff and residents to participate (see the aspect of space below). There are examples of projects in which co-creation with healthcare has worked rather well. For instance, in the project "Onderhuids" (2023–2024), three artists developed art works in co-creation with nurses, in order to reconsider the professional identity of nurses in different nursing contexts, such as community nursing or dementia care departments. Similarly, in "In Search of Stories" (2018–2023), artists worked with terminal cancer patients on co-creating artistic works that reflect on the patients' experiences with the disease, in particular of the diagnosis and how this drastically impacted the understanding of their life stories. However, as these projects held only a loose connection to the healthcare institutions, co-creative processes were naturally much easier to facilitate. Working co-creatively inside of an institution with daily time pressure and high workload on the side of staff remains a challenge.

In hindsight, it seems to us that the artists a) spent too much time with the first exploratory steps: to "find their place" in the institution, so to speak, to get into meaningful contact and exchange with the people there (including constantly changing staff with different schedules); and b) spent this time observing and holding conversations, largely based on spoken language, rather than immediately translating

these experiences into first sketches or small experiments. This resulted in a quite late start of working with and creating artistic materials — a late start of engaging in the (co-)creative and artistic process.

As a more general issue resulting from the initial design of the project, we observed in the staff an expectation of a functional, instrumental stance or expectation towards the arts, — the arts as a problem-solver — and realized that the artists tended to behave towards caring-for, and easily adopted such an instrumental position. In the next few sections, we have compiled a kind of collection or "catalogue" of issues and challenges that emerged during the process of the residencies — starting with exactly the point of instrumentality and function.

*Equality and reciprocity as a point of departure — what is the "function" of art?*

The project and funding application were initiated by Topaz. This, however unintended, somewhat higher degree of influence (which included posing the initial research question up to the phase of reporting) made it challenging for all parties to find a reciprocal balance: We realized an overall tendency to consider the artistic process and work as "helping healthcare professionals," which in our view limited the actual strength and potential of artistic practice. The artists felt they needed to justify their presence in the departments, instead of fully engaging in an artistic and creative process. They experienced that it was much easier to make connections with the residents than with staff.

Based on our experience, in this project as well as earlier ones on the intersection of arts and healthcare, we argue for two important points: First, this work would benefit from a more open, equal, and reciprocal relationship between artist, healthcare professional, and patient/resident/participant. Second, the notion of *not knowing*, while challenging, should be granted a much bigger place in all stages, from observing, creating, and reflecting up to the final report. We witnessed a tendency to aim for concrete, directly "usable" output that can be instrumentalized: "Designing a plant is nothing new, what do we gain with this at Topaz?" In our experience, this tendency results in a loss of the actual stories, the moments in which people were truly moved in their emotions or imaginations, and the reflection on the deeper, somewhat hidden, values underlying the research question. Interestingly, we saw that a new space for ideas, emotions, and perspectives arose in situations when artistic material was part of or impulse for starting an exchange, from where a more connective relationship emerged.

## The curation of the artists

This project made very clear that — as expected and as well-known — social contexts are complex and vulnerable, and healthcare contexts are no different. While it was part of the concept and design of the project to invite young artists, alumni from our institutions, in hindsight we should have spent more time thinking about the exact criteria for choice and curation of artists. This includes criteria such as artistic profile, age, and experience.

While the project was designed as a trajectory including a co-creative artistic approach, it appeared that two of the three artists had insufficient experience with working co-creatively, especially in complex societal contexts. This was not clear enough during the recruitment period. This lack of experience very likely brought them towards "falling back" into the approach of working autonomously in the artist's studio, rather than making their work as much as possible in context, co-creatively. Another aspect is that more experienced artists (more experienced in their own work, in working co-creatively, and with working in social contexts) are typically better able to handle the various, complex parameters of a social context, are more aware of the pitfalls, and can act on them as soon as they appear. In this project, this was delayed, because issues were acknowledged late and supervisors needed to address them. This is no judgement of the three artists participating in this project: The curation was the responsibility of us, the project's core team, and the issue of recruiting young and (in some respects) inexperienced artists was an explicit learning point for us.

In earlier, successful projects of transdisciplinary co-creation between arts and healthcare, we observe the crucial difference that artists were more experienced professionals, with both a well-developed co-creative practice,

as well as sufficient experience in working transdisciplinary and in complex situations and contexts. An example can be found in the project IYANTWAY (*If you are not here, where are you?*) in which artists worked in co-creation with patients experiencing "absences," a light form of epilepsy (Dörr & Hübner, 2017). This project at Topaz provided us with a clear indication that this is an aspect that should not be underestimated or overlooked.

### Was the "assignment" too confusing, unclear, or too open?

At several points during the process we observed a certain confusion among the artists. As supervising team, we wondered if the "assignment," or question to the artists, was actually working in the way we envisioned it from the outset. Did we ask the artists to solve anything, or to engage in an open, co-creative process in a societal context? Should the final products lead to a solution, service, or experience that would improve the target group in one way or another? Does it need to be user-friendly? Could it offer alternative, less expected ideas and questions, or would that already be "too much art?" All three artists were confused regarding responses to these questions, and it remained unclear for them what was expected.

The terminology in the project description did not really help to clarify this, as (in the Dutch context, at least) the term "design" typically implies some kind of problem-solving stance. This implication would also suggest more proximity towards the functional use of art, mentioned above. In supervision conversations we instead aimed for a more open artistic exploration, which produced confusion during the artists' day-to-day work in the departments.

In other words, calling our artists "designers" but treating them as artists caused confusion. We are mentioning this explicitly as this seems to be such a small detail (at least it seemed to us), but it had more substantial consequences than we expected. While we as supervisors and creators of the project might not make such a huge difference between the two terms (or at least understand the "designer-artist"-pair as a flexible relationship in which many hybrid positionalities are possible), others might actually differentiate them, including the artists in residence themselves.

A second point of confusion lay in the initial setup of the project and research question. The artists were asked to look at meaningful relations of the residents before moving to the care institution and how these relationships could be kept intact. However, virtually all of the work happened within the walls of Topaz, while contact with family and friends remained difficult. For residents, the period before moving was already quite far away for them, and questions about this were unproductive. It might have been advisable to change the research question towards what makes life meaningful *in this moment*, without referring to earlier relations too explicitly. Another possibility might have been to include families and friends much more actively from the outset and the design phase, to ensure participation.

Next to this confusion, we wonder if the process we designed worked for the artists. We offered the four stages to them that we mentioned above under "Getting started." However, the artists experienced this approach and process as being too open — and thus as too vague or unclear. This might be one of the reasons they started quite late with creating actual artistic material — an aspect that will come back later in the section "On (too much) reflection through language." This issue also has a direct relationship to what we described in the previous section on curation: While the younger, less experienced artists of our project experienced this process as too open, artists with more experience in this kind of work may have had no issue with openness. For them, it might have provided a more immediate way of working with their own approach, without waiting for clarifying instructions.

### Balance between the artistic and the urge to help

We clearly experienced the challenge to balance the more imaginative and associative artistic process on the one hand, and the urge to "help," to make a meaningful (and instrumental) contribution to what happens in the institution on the other. The initial idea was that the artists would get to know the atmosphere, dynamics, and rhythm of the departments by helping in the day-to-day

routine, and that this would also help them to get into contact with staff and residents. For some of the artists, this was natural to do, while for others this was less the case. However, all artists felt some kind of pressure to help their department, which took time and attention away from their artistic process. In hindsight, we analyze that the time for observing and helping and the time for the artistic process were not balanced; one could argue that the approach "mispositioned" the artists in the departments, exactly because they were immediately instrumentalized.

For the artists, it could be hard to differentiate between personal engagement and the role of a professional. The boundary between the two is personal and different for everyone, and not every artist is willing to step into such a helping role. Especially for the young artists in this project, it proved difficult to guard this boundary.

All three artists were struggling with this "task to help" and its effect on their function and professional identity as artists. Additionally, they felt little space to inquire what actually would have been an appropriate way to interpret "helping" in service of the artistic process. Our analysis is that, as core team of the project, we could or should have spent more time to communicate this particular tension with the organization and staff, in order to create space for the artists to explore this.

### A lack of artistic and safe space

The project showed us how important it is to create a safe space (both physically and mentally) in which artistic processes and dialogues can come to fruition — for everyone involved. Neither us, nor Topaz considered a place such as a studio or other kind of dedicated physical space in which the artists could work. Nor was there any space available for artists to leave materials or objects behind, which posed logistical challenges. Every activity needed to happen surrounded by staff and residents. This led the artists to feel insufficiently free to experiment with materials and methods and hindered them to "step out" of the daily dynamics of the care institution. As soon as they were on site, it was virtually impossible to take more distance and time to think and reflect on their observations. This led to an experienced lack of mental space as well. The artists experienced it as challenging that they repeatedly had to explain that creating artistic work inevitably needs to include preparatory work and research. They felt cramped and de-focused being asked time and time again if they had already "made" anything. In hindsight, a simple space with a door that could be closed would have been a relatively easy solution to help focusing, and to create a sort of "in-between-space", a space safe enough to be a brave space (Cairo et al., 2021, pp. 197-201), into which the artists could invite staff and residents for more artistically-focused explorations and conversations. The artists could have used such a space as their "own world" into which they could invite others and share their artistic process and thinking, on their own ground, so to speak. At the same time, such a space would have brought our concept of a socially engaged residency in a healthcare context closer to the traditional residency form, which offers artists explicitly the space and time to create, without the concern for unhelpful distractions.

### Misguided approach? On (too much) reflection through language

Looking at the overall structure of the process in time, we come to two crucial insights: First, the artists spent a considerable amount of time experiencing Topaz through observation and participation, rather than through creating and making sense of it through artistic creation or experiments. Second, the exchange between artists and the actors in the institution was largely dominated by spoken language, rather than through artistic materials (which are often *not* lingual in nature). And while the institution's staff members were typically highly thankful for the fresh and sometimes unexpected or unorthodox views of the artists, we observed that this exchange through language also had the potential to lead to judgments and unnecessary "tips" towards the institution, which in turn led to defensive behavior of the staff and unhelpful tensions especially between artists and staff. On the other hand, in moments when an artist shared artistic material, sketches, or creative assignments or exercises, the conversation took a much more exploratory, associative, and fruitful direction,

where a true, non-judgmental exchange of ideas took place.

We argue that it is (or actually it should be) exactly the more open, speculative, and creative nature of artistic work and practice that creates a less judgmental and more meaningful dialogue and connection between the arts and institutional healthcare. Artistic materials also have more means and potential to make radical propositions and speculate with them. Statements or thoughts such as "It is only natural that the elderly in such institutions are not happy, as no one asks them what they want[8]" sound unqualified and blunt when spoken, but (after discussion and gaining nuance) can turn into powerful incentives for making artistic work. In turn, this work can facilitate a richer, more empathetic, reciprocal, and impactful dialogue. Taking this further, it is exactly the open, exploratory, and speculative nature of artistic work that can facilitate connectivity, rather than dialogue through spoken language *per se*. It is clear that we have only been able to see this in hindsight, but did not realize the nature of the conversations as being too much focused on language in time. Or — to be more precise — we recognized this, but were unable to change the nature of the conversations accordingly.

An important lesson that has been mentioned a few times already: It is crucial in such residencies that artists start by making, creating material, trying out — just starting, in fact — much earlier, or even as early as possible. This has two important consequences, on the level of documenting experiences and insights, and on the level of reporting.

*On documenting:*
There have been a number of valuable dialogues and encounters between the artists and especially staff members, but we have not been able to document these in a way that is experienceable in other means than written language. To provide an example: In some cases the artists experienced being intensely emotional or sad when they witnessed how the elderly people were not allowed to make specific decisions themselves anymore. While this can be captured in words to some extent (which the artists did, in the form of notes, bullet points or quotes), reading these notes feels pale or shallow compared to the actual experience; it is difficult to understand the true impact and essence of these experiences. If the artists would have been able to capture this in the form of artistic materials (drawings, animations, sounds), this would likely have had the potential to offer a more experiential, deeper and richer way of sharing these experiences.

*On losing insights and nuances in reporting:*
This loss of the quality and nature of the experience directly translates to the final documentation and report of the residencies and the project as a whole. The report has been written by Topaz, the healthcare institution, exclusively in text. The "artistic qualities" (in observations and experiences, concerning ethical, societal or political insights) were almost entirely lost, due to the sole focus on written language and the absence of artistic materials. While the potential of such a project lies in the possibility to come to new insights in healthcare through artistic processes and works, leaving exactly those modes of artistic practice out of a report inevitably results in a loss of this potential — and the imagined and ambitioned innovation in healthcare.

## Final reflections

There are a few aspects that seem helpful to us to be reminded of when it comes to artistic research residencies in healthcare institutions. Obviously, a number of issues and challenges described in the previous sections are entangled: For example, the fact that the artists, being less experienced in co-creation *in situ*, fell back into working in their studio, likely could have been mitigated to some degree (or potentially solved) by having a dedicated space for them in the institution. In summary, we propose a set of five recommendations, which are elaborated on further below:

1. Take sufficient, probably more time than one estimates necessary, to carefully design such a transdisciplinary project, and consider which kinds of preparation everyone needs,
2. Allow artists to take a position of being

[8]This quote is a statement, or rather provocation, uttered by one of the artists during an in-between reflection session with the core team and staff from Topaz.

"function-less" and practice to see this as a meaningful value,
3. Design this "function-less-ness" specifically into a project in the form of time and space for the artistic-creative process,
4. Make sure that artists start working with their respective media and disciplines, and making artistic materials, as early in the process as possible,
5. If possible, provide some kind of dedicated studio space for artists to work in and develop (unfinished) material.

We have learned that taking (even more) time to design such a project carefully is a key factor. This includes the complex task of clarifying and negotiating the expectations for all participating actors: the care institution, its staff, residents, potentially loved ones, artists, and supervising research institutions. These are all participating voices that need a certain amount of synchronization (see Christophe, 2017). This synchronization also requires careful consideration of the relationship between a healthcare institution's question and the approach with which the project will be conducted in and through artistic practice.

If there is one underlying red thread in the challenges we describe, it is the question of what function artists can have in a context of healthcare. Or, more provocatively: In which way can *artists be allowed to be function-less*? Everyone in a healthcare context has a function, including residents, patients, friends, and family. In their best moments, the artists in our residencies experienced themselves as the only ones who are "free of function" (R. Krijgsman, personal communication, November 28, 2023). Resonating with this, Mandy reported being asked: "What are you actually doing here?" We argue that this notion of being "without function" or being "functionless" offers an essential value: to let the arts and artists in healthcare contexts actually come to fruition — which, at the same time, can be confronting and feel risky for healthcare staff, as it works to much against notions of efficiency, clear task-division, and scheduling, which are so ubiquitous in this field. Understandably, being functionless provokes initial reactions that there won't be any (positive) impact or change. However, we argue that, in potential, the exact opposite is the case: It is exactly by facilitating artists with an open environment to connect to a social context that they can find ways that resonate with the underlying questions and issues of this context and develop works, interventions, and artistic utterances that react in a powerful, imaginative, and speculative way to this context — in ways that one could never expect by giving a more or less exact or concrete assignment.

Our recommendation is thus to spend more time thinking about, and designing this function-less-ness[9] into such projects from the outset: to provide more emotional-creative capacity for making, more openness for artists to create and react to what they experience in the context (of a care institution, in this case), and, by this, enable artists to think through their practice — to think through art (Hübner, 2022, p. 8). This is considerably different from the more romantic notions of

## Original purpose

The aim of the collaboration between Topaz and the three artists was to develop new perspectives on the lives of the elderly with their loved ones during and after the move from their home to a care institution. At the same time, we wanted to collect and explore new perspectives on co-creative working in and through the arts in the field of healthcare — through the methodological approach of artistic residencies. We wanted to research the position of the artists in the context and at the same time learn about their artistic skills and talent. Taking this further, we were curious if the open, exploratory, and speculative nature of artistic work was able to facilitate connectivity better than dialogue through spoken language. To create free space for this we wanted to prevent the artists from being positioned or seen as an instrumental "problem solver" in the department, which could limit their artistic freedom. Finally, we were curious about the working or not-working elements of this crossover practice.

[9] Obviously, this is more than just a "negative" argument: Artists can open up, reflect on experiences and issues they encounter, and thus put these "on the agenda", which is not literally without function, but not being "in service" either.

independency, "autonomous art" or "artistic freedom;" we argue for an open sense of situated relationality, where artists can productively relate in a meaningful way, while not being instrumentalized.

Following up on this point, this means that artistic and research practice need to take a much more prominent place, both in terms of literally creating artistic materials, as well as documenting and reflecting along the way. We argue that artists should aim for working with their own media (objects, color, sound, still and moving images, sketches, drawings, and so on) as early as possible. This does not mean they need to create artistic *works* immediately, but rather observe, document, and reflect on a given social context by means of artistic media, which in turn will lead to a less forced but more natural way of creating actual works.

Finally, and again emerging from the previous point, we will reiterate the need for a physical space for artists in such residencies. As mentioned before, a physical space can enhance a clear positioning of artists in an organization, and it can provide them with a "temporary home," both for themselves and for how staff experiences them. A (temporary) studio can be a space to be able to work, where materials can stay, lie around, where artists can play with ideas without having to justify them at any given moment (which can feel unsafe) — and a place to invite staff and residents into. A space safe enough to be brave, and thus a space to make it possible to be an artist in healthcare; not in literal service, not as a therapist, not "in function," but as an artist.

## References

Badura, J., Dubach, S., Haarmann, A., Mersh, D., Rey, A., Schenker, C., & Toro Pérez, G. (2015). *Künstlerische Forschung: Ein Handbuch*. Diaphanes.

Cairo, A., Misiedjan, D., & van Uden, J. (2021). *Holding space. A storytelling approach to trampling diversity and inclusion*. Aminata Cairo Consultancy.

Christophe, N. (2017). The art is in the encounter. In H. Dörr & F. Hübner (Eds.), *If you are not there, where are you?* (pp. 88-105). International Theatre & Film Books Publishers.

Coumans, A. (2020). Ontwerpen in het hier en nu. De artistieke attitude in de zorg voor mensen met dementie. *Forum+, 27*(2), 3-13. https://doi.org/10.5117/FORUM2020.2.002.COUM

Dörr H., & Hübner, F. (Eds.) (2017). *If you are not there, where are you?* International Theatre & Film Books Publishers.

Hübner, F. (2022, November 18). *In good company. Think we must*. Fontys Fine and Performing Arts. https://www.researchcatalogue.net/view/2606237/2606238

Hübner, F. (2024). *Method, methodology and research strategy in artistic research. Between solid routes and emergent pathways*. Routledge.

Madden, R. (2017). *Being ethnographic: A guide to the theory and practice of ethnography*. SAGE Publications Ltd. https://doi.org/10.4135/9781529716689

**Correction notice**
Incorrect Special Issue Labeling (Article erroneously excluded): This article was previously not labeled as part of a special issue due to an error. This has now been corrected.

# Medical Expert Endorsement Fails to Reduce Vaccine Hesitancy in U.K. Residents

Folco Panizza[1], Piero Ronzani[2], Carlo Martini[3,4], Lucia Savadori[5], Matteo Motterlini[3]

[1]Molecular Mind Laboratory, IMT School for Advanced Studies Lucca, Italy
[2]International Security and Development Center, Berlin, Germany
[3]Centre for Applied and Experimental Epistemology, Department of Philosophy, Vita-Salute San Raffaele University, Milan, Italy.
[4]Centre for Philosophy of Social Science, De- partment of Political and Economic Studies, University of Helsinki, Helsinki, Finland.
[5]Cognitive and Experimental Economics Laboratory, Department of Economics and Management, University of Trento, Trento, Italy

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
March 20, 2023
**Accepted**
July 6, 2023
**Published**
September 7, 2023

**Correspondence**
Molecular Mind Laboratory
folco.panizza@imtlucca.it



In this report we outline the null findings of a pre-registered experiment on vaccine hesitancy in the United Kingdom. The experiment targeted vaccine misconceptions common among participants by presenting a correction to such claims endorsed by a group of medical experts. The experiment had the aim to increase vaccination intention and actual uptake during the 2021 COVID-19 vaccination campaign. Our results revealed that, contrary to a similar study conducted with Italian residents, our intervention was unsuccessful in changing participants' attitudes and behaviour towards COVID-19 vaccines. The report concludes with a discussion of the potential reasons for these null findings.

**Keywords** *Expert endorsement, Vaccine hesitancy, Nudging, Debunkning, COVID-19*

Scientists and medical experts are among the professionals trusted the most (Skinner & Clemence, 2022). Are they also the most suitable figures to convince the general public to get vaccinated? In a preregistered experiment, we tested whether expert endorsement increases the effectiveness of debunking messages about COVID-19 vaccines.

Recent literature underscores the significance of debunking in combating misinformation (Lewandowsky et al., 2020). Debunking involves presenting information that directly confronts the core misconception within a particular piece of false news. For instance, if an individual is averse to a vaccine due to a misplaced fear that it could cause autism in children (a case of false information), debunking would confront this belief directly by providing evidence from scientific studies that demonstrate no correlation between vaccination and autism in children. Corrections from experts can be an effective framing and communication strategy to steer people towards recommended policies (Bogliacino et al., 2021). We wanted to test the hypothesis that expert endorsement is an effective intervention to increase positive attitudes towards COVID-19 vaccines and intention to vaccinate.

The design of our debunking intervention was primarily guided by established dual process persuasion theories, such as the elaboration likelihood model (Petty & Cacioppo, 1986), or the heuristic–systematic model (Chaiken, 1980). These theories propose that the credibility of the source, which includes perceived expertise, acts as a persuasive factor: individuals are more likely to be persuaded by a message originating from a credible source than from a less credible one (e.g., Heesacker et al., 1983). This is particularly true when the message is conveyed through a heuristic-peripheral route as opposed to a central-systematic one. Source expertise can significantly contribute to behavior change: debunking based on source credibility takes advantage of heuristic-peripheral processing in a setting (experimental survey) where respondents do not necessarily engage in central processing of information. For this reason, all debunking messages were endorsed by a source that is held in high regard by most of the general population, namely, medical experts and researchers. Interventions also draw upon the extensive body of literature emphasizing the

## Take-home message

An intervention based on medical expert endorsement may not have been successful in improving vaccination intention and actual uptake during the 2021 COVID-19 immunisation campaign in the United Kingdom. The message campaign that was specifically built on experts' advice did not change participants' views or behaviour concerning COVID-19 vaccinations. Further research is needed to determine why a similar intervention succeeded in an Italian sample but not among respondents in the United Kingdom.

role of social norms in changing human behavior (Reynolds, 2019). From this perspective, individuals are perceived as social beings who endeavor to maintain their place in groups composed of individuals they respect, admire, and identify with. In this sense, the opinion of experts influences certain behaviors because experts are seen as a generally respected group of individuals who possess relevant knowledge.

We sought to apply this framework by framing expert endorsement as a social norm: specifically, expert endorsement was presented as a majority consensus of qualified and trusted experts (van der Linden, 2021).

Finally, the interventions were specifically tailored to the sample: messages sent to participants targeted their personal concerns about COVID-19 vaccination as expressed in a prescreening survey. This ensured that the messages were relevant to respondents and increased the likelihood that they would attend the messages.

We monitored a sample of 2,247 people in the United Kingdom through a longitudinal study along the salient phases of the vaccination campaign. Participants in the "expert endorsement" treatment received a series of messages targeting concerns expressed by participants themselves about COVID-19 vaccines. Messages were endorsed by a majority of medical experts consulted on these concerns. In order to minimise demand effects, we collected participants' responses about ten days after the previous debunking message. To test the effectiveness of the intervention, we also monitored beliefs, intentions, and vaccination behaviour of a control group. Contrary to pre-registered hypotheses, vaccination turnout did not increase in the experimental sample compared to control, nor did participants express a higher intention to vaccinate, or more positive beliefs about the protective benefits of vaccines. This lack of evidence contrasts with the results of a similar experiment conducted among Italian residents, (Ronzani et al., 2022) in which the intervention supported by experts was compared with the same intervention supported by a generic group of survey respondents. In the Italian experiment, vaccination intentions and beliefs about the protective benefits of vaccines increased in the expert treatment compared to the non-expert version. Conversely, the same expert intervention had no effect on these measures in the sample of UK participants.[1]

## Methods

Participants were first recruited from the online platform Prolific through a screening survey at the beginning of the vaccination campaign ($N$ = 2598). Collection started on the 12th of January 2021. The goal of this survey was to collect preliminary information about participants' demographics, their initial willingness to receive a vaccine (vaccination intention) and, for vaccine hesitants, their main concern keeping them from getting vaccinated. Questions were adapted from previous Ipsos surveys (Boyon & Silverstein, 2021). Although we did not aim to collect a sample that was representative of the general U.K. population nor did we have any specific demographics predictions, we tried as much as possible to balance the composition of the sample in terms of education. Educational stratification was introduced to match national rates as closely as possible and to avoid biasing the results by, for example, oversampling the educated. Our sample was thus recruited based on quotas defined by the most recent data available about the level of education of U.K. residents (Office for National Statistics, 2019)[2].

[1] For more context about the experimental literature, please consult the twin publication of this report:doi.org/10.1016/j.vaccine.2022.06.031.

[2] We decided to exclude the 'other education' category from the survey since it was not possible to match this category with data available for participants on

**Table 1** Number of participants and retention rate for each wave of the study.

| | Group | |
|---|---|---|
| **Wave** | **Control** | **Experimental** |
| **1**: *April 6–15* | 1063 (100%) | 1057 (100%) |
| **2**: *April 16–25* | 1037 (97.6%) | 1051 (99.4%) |
| **3**: *April 26–May 5* | 1038 (97.7%) | 1046 (99.0%) |
| **4**: *May 6–15* | 1020 (96.0%) | 1004 (95.0%) |
| **5**: *May 16–25* | 983 (92.5%) | 996 (94.2%) |
| **6**: *May 26–June 4* | 957 (90.0%) | 977 (92.4%) |
| **7**: *June 5–14* | 985 (92.7%) | 1010 (95.6%) |

The size of the sample was determined based on the number of available participants in the least represented education category on prolific ("no formal education"). To be eligible, participants must not have received the vaccine at the time of the survey. We tried to collect as many participants as possible given the constraints of the recruiting platform, the goal of having a fairly representative national sample in terms of education level, and the progress of vaccination campaign. We recruited an initial sample of 2598 U.K. residents based on these criteria. Participants were then randomised into an experimental group and a control group, while keeping the proportion of vaccine hesitancy and education balanced between the two groups. 59 participants were excluded in the process because they were missing demographic information (employment data) or because they reported being already vaccinated. We were thus left with a sample of 2539 eligible participants. The Research Ethics Committee of the University of Trento approved the study (protocol no. 2021-001) and subjects provided written informed consent prior to their inclusion. All participants were paid for their time.

The experiment was organised in seven consecutive waves spanning 10 days each. Data collection for the experiment started on the 6th of April 2021. Participants had 10 days to respond to the survey, after which data collection for that wave was closed and a new wave started at the eleventh day. All surveys were scheduled to start at around the same time (14.00 GMT). The longitudinal design of the study included six interventions and a final survey, for a total of seven waves. Although analyses were conducted on participants having completed all seven waves, we also conducted robustness analyses on participants who completed fewer waves. For this reason, we allowed participants to respond to surveys even if they missed previous waves, including the first one[3]. We excluded one participant that moved their residence outside the United Kingdom during data collection. We also excluded one more participant who was missing educational information. We also excluded single responses under specific circumstances. Some participants responded more than once in the same wave, hence we decided to keep their first response only, as the subsequent ones might have been influenced by previous responses. Furthermore, we excluded participant responses from specific analyses in case their responses were not logically plausible. For instance, we excluded data from participants reverting their vaccination status between waves (from "vaccinated" to "not vaccinated") for analyses concerning vaccination behaviour. The final sample size of participants included in any one analysis was $N = 1119$ for the control group, and $N = 1128$ for the expert endorsement group (total $N = 2247$). Table 1 represents the number of participants in each group after exclusion criteria were applied, and the retention rate compared to the initial sample (Supplementary Tables A.3 and A.4 show the same data broken down by level of education; Supplementary Table A.5 shows

Prolific. It may in fact have been the case that foreign degrees that the Office for National Statistics considers as 'other education' (www.ukdataservice.ac.ukmedia262853discover_sqb_education_schneider.pdf) were instead reported by the participants as equivalent to UK degrees, thus producing a mismatch in classification. We thus decided to keep the proportions for the other education levels while removing this category.

We did not carry out any analysis including education as we controlled for this variable through sampling and had no pre-registered hypothesis related to it.

[3]This fact is reflected in the attrition rate of Table 1, which can also be negative (see the increase of participants in the control treatment between wave 2 and wave 3). Robustness analyses include participants who completed fewer waves, with and without excluding participants re-entering the experiment. These analyses yield the same statistical results as in the main text (see Supplementary Analyses).

instead the proportion of participants in each treatment divided by how many waves were completed).

## Experimental Design

Participants responded to up to seven waves. All waves measured our variables of interest (vaccine behaviour, intention and beliefs), and all waves except the last one included a message intervention. In each wave, participants were first asked about their vaccination status (not offered; offered but not vaccinated; vaccinated), their intention to vaccinate (if not vaccinated: "If a vaccine for COVID-19 was offered to me now, I would get it."; four-point response format from "strongly disagree" to "strongly agree"), and their beliefs about vaccines' protective capabilities for themselves and others (two questions: "My vaccination against COVID-19 protects [myself/others]"; seven-point response format from "completely disagree" to "completely agree"). Questions about vaccination status and intention were adapted from previous Ipsos surveys (Boyon & Silverstein, 2021) to keep results comparable.

After responding to the initial questions, participants observed the message interventions (example in Figure 1). One message was created for each wave, and all messages were built around participants' initial concerns about vaccines, which we collected in the preliminary survey. For example, one of the most common concerns was the fact that vaccines had been developed too quickly; the correlated informative intervention stressed the fact that SARS-CoV-2 vaccines' fast development was possible by cutting most bureaucratic times.

For each wave, participants in the expert endorsement treatment received one message intervention that included three parts: Participants were first asked their opinion about the concern targeted in that wave (e.g., "I think one should be vaccinated even if there may be side effects."; options: Yes/No/Don't know). After expressing their opinion, participants in the expert endorsement group observed a message in response to the concern. This response was based on the evaluation of doctors and

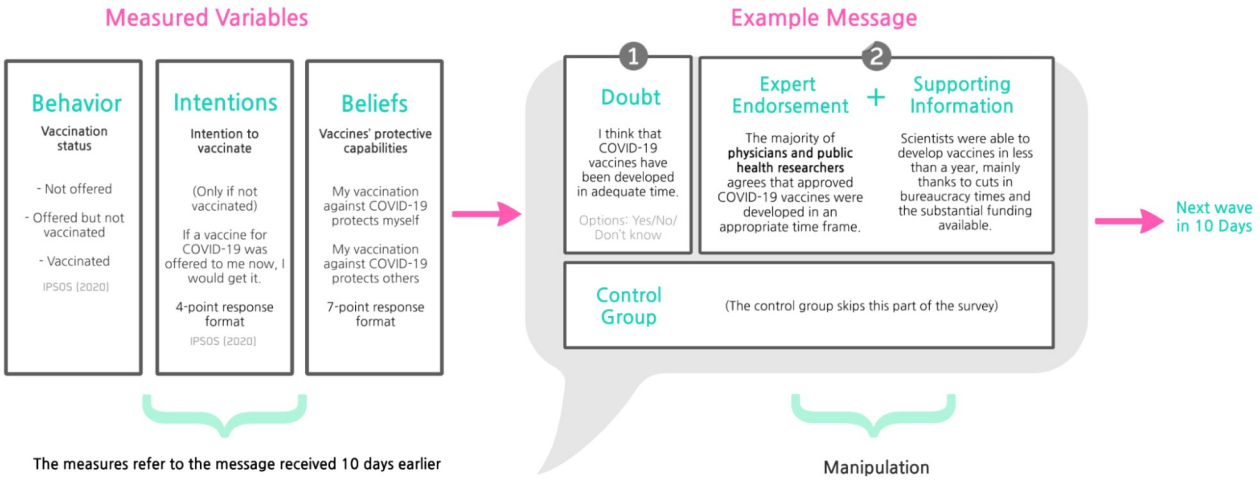



**Figure 1** Flow-chart of one exemplary wave. The wave starts with the recording of vaccination behaviour, intentions, and beliefs (used as outcomes of the previous message intervention). The recording of the measures of interest are followed in the expert treatment group by the debunking message endorsed by experts.

COVID-19 researchers who were shown participants' concerns (see Supplementary Material Expert Survey). The response was phrased as follows: "In the January survey in which you participated, we collected some concerns about vaccination against COVID-19. We recently conducted a second survey among doctors and researchers: The majority of experts agrees that [message]"[4]. The third and last part of the intervention was a text providing support for the endorsement message. Participants in the control treatment did not observe the endorsement nor the message. Messages were based on material from leading health institutions (U.S. Centers for Disease Control and Prevention, European Medicines Agency, U.K. National Health Service). Note that message interventions appeared *after* we collected participants' vaccination status, intention, and beliefs. This ensured that participants' answers were not distorted by any potential demand effect. We expected instead that our messages affected responses in the subsequent wave. The complete list of interventions is available at osf.io/m8cr6.

The last wave did not include a message, but a series of control questions. Participants answered further questions regarding the COVID-19 pandemic and vaccination campaign. Questions included whether participants completed the vaccination cycle and if they contracted COVID-19 in the previous three months (Yes/No questions), whether they would recommend a vaccine to friends and relatives (five-point response format from "completely disagree" to "completely agree"), and a scale measuring coronavirus risk perception (Savadori & Lauriola, 2021). Participants were also asked their main source of information about COVID-19 (multiple choice question) and their trust in the national government, scientists, and pharmaceutical companies ("not at all"/"not much"/"some"/"a lot"/"don't know"; questions adapted from the 2018 Wellcome Global Monitor, 2018). Lastly, participants completed a survey with a series of scales, including the short-form version of the cultural worldview scale (Kahan, 2021), and the Conspiracy Ideation Trait scale (Bode & Vraga, 2018).

[4]The message emphasises that these are not general statistics, but rather data that have been collected by the experimenters from a specific sample of medical researchers. This notwithstanding, we did not include the names or affiliations of the experts interviewed, as this information may have influenced participants' responses more than the message itself, and may have reduced the generalizability of the findings, for example because the name of the expert or institution may have polarised responses.

### Analyses

Analyses were conducted in R (R Core Team, 2018) using the multgee (Touloumis et al., 2013) package package. To test for changes in vaccination uptake we used a Chi squared test comparing the proportion of vaccinated participants between the two experimental treatments. We included only those participants who were offered a dose of vaccine by the end of the experiment ($N$ = 812). To capture changes in vaccination intention, we included only participants who were not yet offered a dose of vaccine at the end of data collection ($N$ = 280). We adopted a repeated measure, ordinal logistic regression for the analysis, including survey wave, experimental group, and their interaction as predictor variables, and participant id as random factor. We interpret the interaction between wave and group as our measure of difference in difference, whereas we consider the non-interaction variables as control measures. Although our analyses focused on participants who completed all waves of the experiment, we include also robustness tests including participants who dropped out before the conclusion of the study and therefore observed fewer messages. Finally, changes in beliefs about vaccines were tested on all participants who completed the study ($N$ = 1599). We adopted the same statistical test as for vaccination intentions, repeated for both our belief questions (protection for self, protection for others). We adopted the 5% significance level to test against the null hypotheses. Post-hoc tests and multiple analyses were corrected for multiple comparisons using a Benjamini-Hochberg procedure. Square brackets indicate family-wise corrected 95% confidence intervals.

## Results

### Vaccination uptake

As part of our pre-registered analyses, we selected all those participants who reported that

they were offered a dose of vaccine between the beginning and the end of the experiment. We tested whether having been assigned to the expert endorsement treatment increased self-reported vaccination uptake compared to control. With the percentage of vaccinated in the last wave being 61.7% in the expert endorsement treatment and 59.6% in the control group, we did not find a significant difference in the percentage of individuals reporting having been vaccinated ($\chi^2(1) = 0.280$, $p = 0.597$, $BF_{10} = 0.103$).

### Intention to vaccinate

Participants' propensity to vaccinate was positively but not significantly affected by expert endorsement, as measured as a difference in difference between experimental and control group across waves (interaction term wave × treatment: $\beta = .008$[-.046, .061], $z = 0.279$, $p = .780$). Instead, vaccination intention increased significantly with time in both groups ($\beta = .042$[.009, .076], $z = 2.501$, $p = .012$). Note that there were no significant differences in intention to vaccinate between the two groups at the beginning of the experiment ($\beta = .123$[-.371, .617], $z = 0.487$, $p = .626$).

Results reported above include only participants who completed the experiment. We additionally explored how many messages are sufficient to observe a significant effect of expert endorsement on intention[5]. Table 2 reports results including different subsets of the sample: the first row includes only participants who read all 6 messages (results above), whereas the last one includes participants who read at least 1 message or more. Regardless of the messages exposed, effect of time is significant and robust, whereas the effect of the intervention remains non-significant.

[5]Interpretation of these analyses is valid if there are no confounding factors affecting how many waves participants completed before dropping out. In other words, whether or not taking part in some waves of the study should not be dictated by endogenous factors. A potential confound is that only participants who were strongly motivated completed multiple consecutive waves. For this reason, we allowed participants to re-enter the experiment even after missing waves. We repeated the analysis by including these data points and found comparable results to the ones reported in the main text (Supplementary Table Appendix A.4).

### Beliefs about vaccines

Regression analyses for vaccine beliefs did not reveal any significant effect of expert endorsement: after the experiment, participants in the experimental group reported around the same beliefs about the protectiveness of vaccines compared to the control group. This was true for both questions, protection to self ($\beta = -.010$[-.031, .012], $z = -0.873$, $p = .383$) and protection for others ($\beta = .009$[-.013, .031], $z = 0.816$, $p = .414$). Our control variables suggest that beliefs about the protection for others did increase over time in both groups ($\beta = .030$[.015, .044], $z = 4.063$, $p < .001$), but this increase was not significant for beliefs about the protection for self ($\beta = .007$[-.008, .022], $z = 0.958$, $p = .338$). Our tests also indicate that beliefs did not significantly differ at the beginning of the experiment (self: $\beta = .066$[-.108, .240], $z = 0.744$, $p = .457$; others: $\beta = .025$[-.146, .195], $z = 0.284$, $p = .776$). As a final robustness check, we test whether the role of the expert is also significant when including dropped-out participants. These tests confirm the non-significant effect of the intervention (Supplementary Tables in Appendix A.5).

## Discussion

This study aimed at testing the effectiveness of an intervention meant to promote positive beliefs about vaccines and to increase vaccination intention and uptake in a sample of U.K. residents. The intervention consisted of providing participants with pieces of information about the COVID-19 vaccine that addressed their reasons for being hesitant in vaccinating (debunking information). Concerns were expressed by the participants themselves in a preliminary survey, and response messages were vetted by a team of medical experts and researchers. Informative text snippets were provided to the same individuals in seven different waves, 10 days apart from each other. To test the effect of the intervention, we also monitored vaccine beliefs, intentions and behaviour in a control group that did not receive any of the messages.

Results show that expert endorsement did not have a significant effect on vaccination uptake, nor on vaccination intentions or beliefs about the protectiveness of COVID-19 vaccines.

**Table 2** Vaccination intention as a function of the number of consecutive messages read.

| | N | Expert endorsement | | | Control group | | | Baseline differences | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Messages | | *β* | z | *p* | *β* | z | *p* | *β* | z | *p* |
| 6 | 280 | 0.008 [-0.046,0.061] | 0.279 | 0.780 | 0.042 [0.009,0.076] | 2.501 | 0.012* | 0.123 [-0.371,0.617] | 0.487 | 0.626 |
| **5+** | 295 | 0.009 [-0.040,0.059] | 0.373 | 0.709 | 0.043 [0.012,0.074] | 2.745 | 0.006** | 0.096 [-0.381,0.573] | 0.395 | 0.693 |
| **4+** | 323 | 0.009 [-0.040,0.057] | 0.347 | 0.729 | 0.043 [0.012,0.074] | 2.752 | 0.006** | 0.044 [-0.407,0.496] | 0.192 | 0.847 |
| **3+** | 355 | 0.007 [-0.042,0.056] | 0.281 | 0.778 | 0.045 [0.013,0.077] | 2.746 | 0.006** | -0.013 [-0.449,0.424] | -0.056 | 0.955 |
| **2+** | 386 | 0.004 [-0.045,0.053] | 0.168 | 0.867 | 0.044 [0.011,0.076] | 2.642 | 0.008** | -0.028 [-0.450,0.395] | -0.129 | 0.897 |
| **1+** | 416 | 0.006 [-0.043,0.055] | 0.227 | 0.820 | 0.042 [0.010,0.074] | 2.574 | 0.010* | -0.010 [-0.419,0.398] | -0.050 | 0.960 |

Our pre-registered analyses yielded null results, thus not supporting the original predictions. These results come in contrast to findings from an Italian sample who underwent a similar intervention (Ronzani et al., 2022). One design deviation from the current study was that the control group also received the intervention messages, but these were endorsed by a generic "majority of respondents" (thus not specifying the qualifications of the experts). In this study, we found that while vaccination uptake did not significantly increase, both vaccination intentions and beliefs were more positive after the intervention.

What factors could explain the differences between these two studies? Below are a number of hypotheses that could in part explain this gap. Firstly, data from the two studies were collected in parallel, but the phases of the vaccination campaigns in the two countries did not coincide. As a reference, half of the eligible Italian population was administered at least a dose of the vaccine by the first week of July 2021, after the start of the experiment. In the UK, this event occurred around mid-March, much earlier than in Italy and before the start of the experiment. Not only was the timing of the campaign different, but also the policies discussed, such as the European COVID certificate, as well as the results of negotiations with vaccine companies. Indeed, since Brexit, many key negotiations and policies have been conducted separately for the UK and EU countries, which in turn may have influenced the topics covered in the media and public opinion. Distinct conditions could partly explain the non-significant results in the United Kingdom: a higher rate of vaccinations at the beginning of the experiment reduced the sample size available for analyses about vaccination intention: indeed, of those who completed all experimental waves, only 12 participants in the expert endorsement treatment initially reported that they were unwilling to be vaccinated when possible, compared to 20 in the control treatment. This may have contributed to a ceiling effect in the effectiveness of the intervention. However, a small number of vaccine hesitant participants would still not explain the non-significant difference between the two groups with regard to beliefs about the protectiveness of the vaccine. In fact, the number of sceptics was considerably larger and more balanced among the treatments. In this respect, other elements may have contributed to the non-significant effect of our intervention, such as a different responsiveness to our messages. For example, exploratory analyses (Appendix A.3) suggest that the debunking message displayed in the first wave (about the time frame for vaccine development) was more likely to address concerns expressed by the Italian sample than by UK respondents, potentially making our intervention less effective. These and other differ-

## Original purpose

The study's original aim was to test whether expert endorsement improves the impact of debunking messages about COVID-19 vaccines in the United Kingdom. The intervention targeted common vaccine misunderstandings held by participants by presenting a correction to such beliefs backed up by a panel of medical professionals. The ultimate goal of the study was to increase vaccination intention and actual uptake during the 2021 COVID-19 immunization campaign.

ences may be the result of an effective communication campaign by the U.K. government or the National Healthcare System. Indeed, our data show that over the course of the experiment there was a significant increase in vaccination intentions and beliefs about the protective capabilities of vaccines towards others. Although there is no significant increase in the belief that vaccines can protect oneself, the numbers related to this belief were already quite high at the beginning of the experiment. Additional differences might have contributed to the observed results, such as socio-cultural differences (e.g., the entrenchment and spread of no-vax movements in the two countries, trust in the healthcare system, etc.), the level of education in the two samples (stratified in the U.K. sample, unstratified and generally high in the Italian sample), as well as events of national resonance, such as the suspension of the Astra-Zeneca vaccine in Italy. These various discrepancies make comparing the two data sets an arduous task, despite the similarity of the experimental designs.

One more explanation to our non-significant results that seems unlikely is experimenter demand effect (Zizzo, 2009). This effect predicts that participants report differently from their real intentions because they want to fulfil the experimenter's presumed expectations. Participants in the control treatment may have conveniently concealed the offer of vaccination in order not to report their refusal of the vaccine, or they may have declared their intention to vaccinate while remaining, in reality, hesitant. We remain unconvinced by this explanation, as the control treatment did not cue any kind of intention on the part of the experimenters, thus making it unlikely that the participants unambiguously changed their behaviour towards a more pro-vaccine attitude. One final explanation is that the intervention that we designed might simply not be effective in certain populations (Bryan et al., 2021). As we note in Ronzani et al. (2022) The cultural characteristics of Italy make it peculiar in more than one way, and thus we may observe a certain degree of heterogeneity between populations more or less similar to this specific country. Further replications of the current design in different regions of the world will be needed to verify this explanation.

## Conclusions

In this longitudinal study that followed a group of U.K. residents over the course of several months, we recorded their concerns, beliefs, attitudes and choices about vaccines. By offering information endorsed by experts addressing the main doubts raised by hesitant people (debunking), we attempted to increase participants' intention to vaccinate and, consequently, vaccination uptake. Our results, however, reveal that our intervention was ineffective in achieving these results, or in changing participants' beliefs about the protectiveness of vaccines. Further research is needed to understand why a similar intervention has worked in an Italian context but not among residents of the United Kingdom.

## Acknowledgments

We would like to thank Sergiu Burlacu and Austeja Kazemekaityte for advice on the design and Maria Almudena Claassen for her magic plot advice. We would also thank the attendee of SPUDUM 2021, TIBER Symposium 2021, and INEM 2021.

## Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870883. The information and opinions are those of the authors and do not necessarily reflect the opinion of the European Commission.

## References

Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication, 33*(9), 1131–40. https://doi.org/10.1080/10410236.2017.1331312

Bogliacino, F., Charris, R., Gómez, C., Montealegre, F., & Codagnone, C. (2021). Expert endorsement and the legitimacy of public policy. Evidence from Covid19 mitigation strategies. *Journal of Risk Research, 24*(3-4), 394–415. https://doi.org/10.31235/osf.io/zbqjd

Boyon, N., & Silverstein, K. (2021). Global attitudes: COVID-19 vaccines. *Ipsos*. https:

//www.ipsos.com/en-ro/global-attitudes-covid-19-vaccine-january-2021

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour, 5*(8), 980–9. https://doi.org/10.1038/s41562-021-01143-3

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Heesacker, M., Petty, R. E., & Cacioppo, J. T. (1983). Field dependence and attitude change: Source credibility can alter persuasion by affecting message-relevant thinking. *Journal of Personality, 51*(4), 653–66. https://doi.org/10.1111/j.1467-6494.1983.tb00872.x

Kahan, D. M. (2021). Cultural cognition as a conception of the cultural theory of risk. In S. Roeser, R. Hillerbrand, P. Sandin, & M. Peterson (Eds.), *Handbook of risk theory: Epistemology, decision theory, ethics, and social implications of risk*. Springer Dordrecht. https://doi.org/10.1007/978-94-007-1433-5_1

Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarracín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E., Rand, D. G., N., R. D., J., R., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M., Swire-Thompson, B., van der Linden, S., & Vraga, E. K. (2020). Databrary. https://doi.org/10.17910/b7.1182

Office for National Statistics. (2019). Highest level of qualification achieved by people living in UK regions. https:/ / www .ons.gov.uk/ peoplepopulationandcommunity/ educationandchildcare/ adhocs/10516highestlevelofqualificationachievedbypeoplelivinginukregions2010to2018

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. Springer. https://doi.org/10.1007/978-1-4612-4964-1

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reynolds, K. J. (2019). Social norms and how they impact behaviour. *Nature Human Behaviour, 3*(1), 14–15. https://doi.org/10.1038/s41562-018-0498-x

Ronzani, P., Panizza, F., Martini, C., Savadori, L., & Motterlini, M. (2022). Countering vaccine hesitancy through medical expert endorsement. *Vaccine, 40*(32), 4635–43.

Savadori, L., & Lauriola, M. (2021). Risk perception and protective behaviors during the rise of the COVID-19 outbreak in italy. *Frontiers in Psychology, 11, Article 577331*.

Skinner, G., & Clemence, M. (2022). Ipsos veracity index 2022. *Ipsos*. https://www.ipsos.

Touloumis, A., Agresti, A., & Kateri, M. (2013). GEE for multinomial responses using a local odds ratios parameterization. *Biometrics, 69*(3), 633–40. https://doi.org/10.1016/j.vaccine.2022.06.031

van der Linden, S. (2021). The gateway belief model (GBM): A review and research agenda for communicating the scientific consensus on climate change. *Current Opinion in Psychology, 42*, 7–12. https://doi.org/10.1016/j.copsyc.2021.02.005

Zizzo, D. J. (2009). Experimenter demand effects in economic experiments. *Experimental Economics, 13*(1), 75–98. https://doi.org/10.2139/ssrn.1163863

## Appendix A. Materials and Methods

### Appendix A.1. Waves completed

**Table A.3** Number of participants and retention rate for each wave of the study, by level of education, Control group.

| | | | | Education | | | |
|---|---|---|---|---|---|---|---|
| **Wave** | no qualifications | secondary | high school | college | undergraduate | graduate | doctorate |
| **1**: *April 6–15* | 55 (100%) | 174 (100%) | 375 (100%) | 106 (100%) | 228 (100%) | 109 (100%) | 16 (100%) |
| **2**: *April 16–25* | 52 (94.55%) | 177 (101.72%) | 355 (94.67%) | 109 (102.83%) | 220 (96.49%) | 109 (100%) | 15 (93.75%) |
| **3**: *April 26–May 5* | 50 (90.91%) | 174 (100%) | 357 (95.2%) | 109 (102.83%) | 223 (97.81%) | 111 (101.83%) | 14 (87.5%) |
| **4**: *May 6–15* | 51 (92.73%) | 172 (98.85%) | 349 (93.07%) | 110 (103.77%) | 220 (96.49%) | 103 (94.5%) | 15 (93.75%) |
| **5**: *May 16–25* | 46 (83.64%) | 163 (93.68%) | 334 (89.07%) | 108 (101.89%) | 214 (93.86%) | 104 (95.41%) | 14 (87.5%) |
| **6**: *May 26–June 4* | 47 (85.45%) | 156 (89.66%) | 324 (86.4%) | 104 (98.11%) | 214 (93.86%) | 96 (88.07%) | 16 (100%) |
| **7**: *June 5–14* | 51 (92.73%) | 163 (93.68%) | 335 (89.33%) | 106 (100%) | 213 (93.42%) | 102 (93.58%) | 15 (93.75%) |

**Table A.4** Number of participants and retention rate for each wave of the study, by level of education, Expert group.

| | | | | Education | | | |
|---|---|---|---|---|---|---|---|
| **Wave** | no qualifications | secondary | high school | college | undergraduate | graduate | doctorate |
| **1**: *April 6–15* | 49 (100%) | 176 (100%) | 367 (100%) | 102 (100%) | 230 (100%) | 114 (100%) | 19 (100%) |
| **2**: *April 16–25* | 51 (104.08%) | 179 (101.7%) | 360 (98.09%) | 109 (106.86%) | 221 (96.09%) | 116 (101.75%) | 15 (78.95%) |
| **3**: *April 26–May 5* | 51 (104.08%) | 183 (103.98%) | 358 (97.55%) | 105 (102.94%) | 217 (94.35%) | 117 (102.63%) | 15 (78.95%) |
| **4**: *May 6–15* | 48 (97.96%) | 179 (101.7%) | 344 (93.73%) | 102 (100%) | 209 (90.87%) | 107 (93.86%) | 15 (78.95%) |
| **5**: *May 16–25* | 51 (104.08%) | 179 (101.7%) | 335 (91.28%) | 101 (99.02%) | 209 (90.87%) | 105 (92.11%) | 16 (84.21%) |
| **6**: *May 26–June 4* | 48 (97.96%) | 169 (96.02%) | 332 (90.46%) | 99 (97.06%) | 211 (91.74%) | 105 (92.11%) | 13 (68.42%) |
| **7**: *June 5–14* | 52 (106.12%) | 181 (102.84%) | 342 (93.19%) | 103 (100.98%) | 212 (92.17%) | 106 (92.98%) | 14 (73.68%) |

**Table A.5** Proportion of participants in each treatment by number of waves completed. Note: this table includes for reference also participants who completed only one wave, although these participants were not included in the analyses as we could not measure the impact of our message intervention.

| | Group | |
|---|---|---|
| **Waves** | **Expert** | **Control** |
| **7**: | 68.5% | 67.9% |
| **6**: | 10.1% | 10.1% |
| **5**: | 5.2% | 5.3% |
| **4**: | 4.7% | 5.4% |
| **3**: | 4.3% | 3.0% |
| **2**: | 3.2% | 3.9% |
| **1**: | 3.8% | 4.4% |

## Appendix A.2. Expert Survey

After collecting concerns about COVID-19 vaccines from vaccine hesitants in the prescreening survey, we asked COVID-19 researchers to express their agreement on a selection of rebuttals to such doubts. Researchers were recruited through word of mouth at the authors' host institution and related research centres. These experts were asked to fill a short survey where they rated their level of agreement with a series of statements (e.g., "COVID-19 vaccines were developed in an appropriate time frame."; five-point response format, from "Strongly agree" to "Strongly disagree"). We received 10 responses (the number was not disclosed in the experiment), and sorted statements by the level of agreement between respondents. We then selected those claims that received support by a majority of respondents and included them as messages in the experiment (osf.io/m8cr6 for the full list).

## Appendix A.3. Concern differences between samples

In each wave of the experiment, participants in the treatment group and participants in the Italian sample were asked about their agreement with one of the many concerns that they had originally raised in the pre-screening phase. As an exploratory analysis, we compared how many respondents still agreed with these initial doubts, and compared this level of agreement between the two samples. We found that in the first wave, the Italian sample was much more sceptical about the time frame in which the vaccines were developed: only 58% of Italian participants agreed with the statement "I think that COVID-19 vaccines were developed in an appropriate time frame", compared to 71% of UK participants. This difference is statistically significant ($\chi^2(2) = 71$, $p < 0.001$; $BF_{10} = 4.8 \times 10^{13}$). Conversely, Italian respondents were less sceptical than UK respondents in waves 3 ("it is important that all eligible individuals get vaccinated;" 93% versus 88% agreement, $\chi^2(2) = 24$, $p < 0.001$; $BF_{10} = 66.5$) and 6 ("vaccines can protect people from virus mutations;" 65% versus 59% agreement, $\chi^2(2) = 17$, $p < 0.001$; $BF_{10} = 9.8$). However, please note the following two caveats: first, we did not ask these questions in the UK control treatment, so we only have data for half of that sample. Second, our intervention could also have affected changes in agreement in later waves, making concern differences in subsequent waves less obvious to interpret. With these limitations in mind, it is still interesting to observe an initial difference of opinion distinguishing the two samples.

## Appendix A.4. Vaccination intention including non-consecutive participation

**Table A.6** Vaccination intention as a function of the number of messages read (including non-consecutive participation.).

| | N | Expert endorsement | | | Control group | | | Baseline differences | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Messages | | *β* | z | *p* | *β* | z | *p* | *β* | z | *p* |
| **6** | 1599 | -0.010 [-0.031,0.012] | -0.873 | 0.383 | 0.007 [-0.008,0.022] | 0.958 | 0.338 | 0.066 [-0.108,0.240] | 0.744 | 0.457 |
| **5+** | 1635 | -0.009 [-0.030,0.013] | -0.791 | 0.429 | 0.006 [-0.008,0.021] | 0.840 | 0.401 | 0.059 [-0.110,0.228] | 0.688 | 0.491 |
| **4+** | 1728 | -0.009 [-0.030,0.012] | -0.835 | 0.404 | 0.006 [-0.009,0.021] | 0.804 | 0.421 | 0.052 [-0.113,0.218] | 0.623 | 0.533 |
| **3+** | 1806 | -0.009 [-0.030,0.012] | -0.804 | 0.421 | 0.006 [-0.009,0.020] | 0.757 | 0.449 | 0.058 [-0.103,0.219] | 0.702 | 0.483 |
| **2+** | 1893 | -0.010 [-0.030,0.011] | -0.906 | 0.365 | 0.007 [-0.008,0.021] | 0.893 | 0.372 | 0.068 [-0.089,0.226] | 0.849 | 0.396 |
| **1+** | 1973 | -0.009 [-0.029,0.012] | -0.828 | 0.408 | 0.006 [-0.008,0.021] | 0.845 | 0.398 | 0.053 [-0.100,0.206] | 0.676 | 0.499 |

**Table A.7** Protectiveness of vaccine for self: regression results (uncorrected) as a function of the number of messages read.

| | N | Expert endorsement | | | Control group | | | Baseline differences | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Messages | | *β* | z | *p* | *β* | z | *p* | *β* | z | *p* |
| **6** | 280 | 0.008 [-0.046,0.061] | 0.279 | 0.780 | 0.042 [0.009,0.076] | 2.501 | 0.012* | 0.123 [-0.371,0.617] | 0.487 | 0.626 |
| **5+** | 345 | 0.017 [-0.030,0.063] | 0.699 | 0.485 | 0.040 [0.010,0.070] | 2.602 | 0.009** | 0.052 [-0.389,0.494] | 0.232 | 0.817 |
| **4+** | 387 | 0.020 [-0.026,0.066] | 0.844 | 0.399 | 0.039 [0.008,0.070] | 2.495 | 0.013* | -0.035 [-0.459,0.388] | -0.164 | 0.870 |
| **3+** | 434 | 0.025 [-0.022,0.072] | 1.035 | 0.301 | 0.037 [0.007,0.068] | 2.374 | 0.018* | -0.051 [-0.461,0.359] | -0.242 | 0.809 |
| **2+** | 468 | 0.022 [-0.025,0.069] | 0.917 | 0.359 | 0.037 [0.006,0.069] | 2.310 | 0.021* | -0.082 [-0.479,0.315] | -0.407 | 0.684 |
| **1+** | 500 | 0.017 [-0.030,0.063] | 0.703 | 0.482 | 0.039 [0.009,0.068] | 2.555 | 0.011* | 0.042 [-0.340,0.425] | 0.217 | 0.828 |

## Appendix A.5. Protectiveness beliefs as a function of number of messages read

**Table A.8** Protectiveness of vaccine for others: regression results (uncorrected) as a function of the number of messages read.

| | N | Expert endorsement | | | Control group | | | Baseline differences | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Messages | | *β* | z | *p* | *β* | z | *p* | *β* | z | *p* |
| **6** | 1599 | 0.009 [-0.013,0.031] | 0.816 | 0.414 | 0.030 [0.015,0.044] | 4.063 | <0.001*** | 0.025 [-0.146,0.195] | 0.284 | 0.776 |
| **5+** | 1635 | 0.008 [-0.013,0.029] | 0.725 | 0.469 | 0.030 [0.016,0.044] | 4.175 | <0.001*** | 0.015 [-0.151,0.180] | 0.175 | 0.861 |
| **4+** | 1728 | 0.007 [-0.014,0.028] | 0.644 | 0.519 | 0.031 [0.016,0.045] | 4.207 | <0.001*** | 0.011 [-0.151,0.173] | 0.132 | 0.895 |
| **3+** | 1806 | 0.008 [-0.013,0.029] | 0.752 | 0.452 | 0.030 [0.015,0.044] | 4.072 | <0.001*** | -0.006 [-0.164,0.153] | -0.071 | 0.943 |
| **2+** | 1893 | 0.008 [-0.013,0.029] | 0.725 | 0.468 | 0.030 [0.015,0.044] | 4.085 | <0.001*** | 0.000 [-0.155,0.155] | -0.002 | 0.998 |
| **1+** | 1973 | 0.008 [-0.013,0.029] | 0.760 | 0.447 | 0.029 [0.015,0.044] | 4.092 | <0.001*** | -0.007 [-0.157,0.143] | -0.090 | 0.928 |

**Correction notice**
Incorrect Special Issue Labeling (Article erroneously excluded): This article was previously not labeled as part of a special issue due to an error. This has now been corrected.

# On the Significance of Place: Vaccination Refusal as a Situated Phenomenon

Martijn van der Meer[1]

Expert endorsement seems a promising tool in countering vaccine hesitancy. Yet findings from an experiment in the United Kingdom, published in this journal, found that repeated expert backed "debunking" messages had little effect on vaccination intentions or behaviors. At the same time, a similar study in Italy had earlier observed a slight increase in the intent to vaccinate—despite actual uptake remaining unchanged. In this article, I reflect on the differences between these studies and argue for a situated public health approach: one that opens up on diversity and responds to local trust dynamics, cultural nuances, and community values in shaping vaccine responses. Some publics may be reassured by scientific endorsement. Others could remain cautious, influenced by deeply held beliefs about risk and norms and values prioritized within their communities. I propose to interpret the null result published in the *Journal of Trial and Error* as pointing in the direction of public health communication strategies that move beyond a one-size-fits-all model, adapting to the unique social landscapes in which individuals live together and the places where their views are formed and expressed.

**Keywords** *vaccination hesitancy, vaccination refusal, conceptual replication, public health, place, knowledge deficit*

[1]Kavli Center for Ethics, Science, and the Public, University of California, Berkeley, United States
[2]Department of Public Health, Erasmus MC, Rotterdam, the Netherlands
[3]Department of History, Erasmus University, Rotterdam, the Netherlands

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
January 23, 2025
**Accepted**
March 12, 2025
**Published**
April 1, 2025

**Correspondence**
Erasmus University
vandermeer@eshcc.eur.nl



## Introduction

How do we ensure that an entire country gets vaccinated? By "we," I mean scientists and policymakers committed to safeguarding population health. Since nation-states began prioritizing the health of their populations, vaccination has become a deeply political issue (e.g., Porter, 2005; Vargha & Wilkins, 2023). A single jab operates on two levels: It protects the individual from future disease while also subjecting them to potential side effects. At the societal level, widespread vaccination can lead to "herd immunity," shielding the population once enough individuals are immunized. Achieving this state requires healthy individuals to accept the small personal risks of vaccination, even when the likelihood of encountering the pathogen or developing side-effects is negligible.

Yet—or perhaps, so: Not everyone is willing to participate in immunization programs. Some doubt the efficacy of vaccines, even when they are clinically approved and provided by governments. Others prioritize the risks for themselves to outweigh the benefits for the population (Hobson-West, 2003). Whether these concerns are valid or not, refusal presents a challenge for public health authorities: Non-participation undermines the collective goal of population immunity. This is especially problematic in contexts where respect for individual autonomy and bodily integrity prevents governments from mandating vaccination. If vaccination remains a matter of personal choice, scientists and policymakers must grapple with a pressing question: How can we change beliefs, intentions, and, ultimately, behavior to ensure both individual protection and collective health?

Historically speaking, this question is not new. But it has gained new urgency as vaccination rates in national child immunization programs—particularly in the global North—continue to decline. In many countries where

vaccines are available for free or at low cost, a growing number of parents have refused to vaccinate their children against diseases such as measles, whooping cough, rubella, and tetanus (e.g., Flatt et al., 2024; Seither, 2024; van Lier et al., 2024). In response, some ethicists have started tying access to shared spaces such as kindergartens to vaccination status, echoing measures introduced during the COVID-19 pandemic (Pierik & Verweij, 2024). At the height of that public health disaster, multiple national governments employed digital tools to restrict access to public spaces. In doing so, they temporarily transformed their countries into social laboratories, testing how individual behavior could be influenced to achieve collective goals by providing them with better or more information. Social scientists such as Folco Panizza and his colleagues (2023) took up this challenge and came up with an experiment.

As part of a Horizon 2020 project on "Policy, Expertise, and Trust in Action," Panizza and his colleagues set out to investigate whether hesitant individuals could be persuaded when confronted with the consensus of scientists and medical experts (European Commission, 2024). The researchers focused not on the *message* but on the *messenger*. Would explicit expert endorsement change minds? In 2021, the team conducted an experiment targeting vaccine-hesitant individuals and exposed them to repeated "debunking" messages over 3 months. The research ran simultaneously in two countries: Italy and the United Kingdom. Based on the order of listed co-authors, it appears that Piero Ronzani led the Italian arm of the study, while Folco Panizza oversaw the UK component.

It was Piero who got lucky.

In June 2022, *Vaccine*—a Q1 journal in public health and immunology—published the initial findings under the title "Countering vaccine hesitancy through medical expert endorsement." The results from Italy showed that "scientist[s] and medical experts are not simply a generally trustworthy category but also a well[-]suited messenger in contrasting disinformation during vaccination campaigns" (Ronzani et al., 2022, p. 4635). Over a year later, the same research group, now led by Panizza, published a second paper, this time with a starkly different title: "Medical expert endorsement fails to reduce vaccine hesitancy in U.K. residents" (Panizza et al., 2023). A nearly identical experimental design had yielded completely different outcomes. In the United Kingdom, expert endorsement appeared to have no significant impact on vaccination intentions. This divergence raises important questions: Why did the intervention succeed in Italy but fail in the UK? Is the difference simply a matter of replication, or does it reveal something about the contexts in which these studies took place? And if so, then what?

Picking up on these questions, my intention with this reflection article is to contribute to theory building on vaccine hesitancy. I argue that Ronzani and Panizza's work supports the hypothesis that the effectiveness of expert endorsement hinges on localized trust dynamics and the alignment of public health engagement with community priorities. To develop this argument, I first compare the findings from Italy and the UK studies, situating the UK study as a conceptual replication. While direct replication reproduces an experiment as closely as possible to verify its findings, conceptual replication tests the same hypothesis with modifications in design, population, or conditions to assess its generalizability (e.g., Crandall & Sherman, 2016; Schmidt 2009). Beyond this comparison, I draw on recent work in the medical humanities to argue that vaccination campaigns should expand from a singular reliance on the authority of scientific claims toward a more contextualized form of engagement. Finally, I briefly sketch the broader implications of these findings for public health strategies aimed at countering vaccine hesitancy.

## Expert endorsement does not work universally

To test the hypothesis that hesitant citizens might change their minds about vaccination when addressed by medical and scientific experts, the Horizon 2020 research team designed an online experiment. They recruited 2,277 participants in Italy and 2,247 participants in the UK via the Prolific platform (Palan & Schitter, 2018). Following a randomized controlled trial design, participants were divided into experimental and control groups. Over

a 3-month period, the experimental groups received debunking messages explicitly endorsed by experts. For instance, to counter fears about the rapid development of COVID-19 vaccines, participants were reassured that accelerated timelines stemmed from minimizing bureaucratic delays rather than compromising scientific rigor. In both studies, data were collected over seven waves between April and June 2021, spaced 10 days apart. To measure the impact of expert endorsement, the researchers assessed three outcomes: whether participants got vaccinated during the study, their intention to do so, and any changes in beliefs about vaccines' protective capabilities.

The two experiments also differed—not just in terms of an Italian versus a UK study population, but also in subtle aspects of study design. First, the Italian study tailored its debunking messages based on participants' responses in a pre-trial exploratory phase, whereas in the UK, the same materials used for Italian participants were simply re-used without local customization (see Appendix 3 in: Panizza et al., 2023). Second, the control groups were structured differently: In Italy, control group participants received debunking messages endorsed by "generic respondents" (e.g., "the majority of survey participants"), whereas in the UK, the control group received no debunking messages at all. This distinction allowed the Italian study to measure expert endorsement against a baseline of general messaging, while the UK experiment tested the presence or absence of expert-endorsed messaging alone. Third, as the research team themselves note, both experiments were "quasi-experimental," reflecting different real-world contexts. The UK rollout of COVID-19 vaccines was already well underway by the time its study began—meaning most citizens had already received a vaccine—whereas the Italian vaccination campaign had only just began (Panizza et al., 2023).

These differences in design, population, and context make it clear that the two studies are not direct replications. Nevertheless, they both tested the same hypothesis—whether expert-endorsed messaging can reduce vaccine hesitancy—across distinct contexts (Crandall & Sherman, 2016; Schmidt 2009). As such, the UK experiment can still be considered a "conceptual replication" of the Italian study, even given the three most important differences in study design. There are three key reasons why. First, while the Italian study customized its messages based on an exploratory phase, the UK study did not, which may have enhanced effectiveness in Italy but did not alter the goal to assess if expert endorsement causes a shift in vaccine *behavior*. Second, the Italian control group received messages attributed to generic respondents, whereas the UK control group received no debunking messages at all—yet both experiments measured whether expert endorsement affected participants' attitudes *relative to a baseline*. Third, the timing of the UK study may have limited its impact on vaccination intent, as most participants had already received the COVID-19 vaccine by then. However, as the researchers also note, this so-called "ceiling effect" does not explain the lack of change in vaccine-related *beliefs*. If expert-endorsed messaging were universally effective, one would expect it to influence beliefs about vaccine effectiveness, even in a population with high vaccination rates—and that was not what the researchers observed.

Despite testing the same intervention, the two studies produced contrasting results. The Italian study showed a significant effect of expert endorsement on the intention to get vaccinated and on the belief that vaccination protects others. However, there was no significant effect on the belief that vaccination protects oneself, and neither expert endorsement nor tailored debunking led to higher actual vaccination uptake. In contrast, the UK study found that intentions to vaccinate increased in both the experimental and control groups, with no significant difference between them. Moreover, the intervention had no meaningful effect on vaccination behavior or beliefs regarding the protective capabilities of the COVID-19 vaccines—whether for oneself or the community. Only in Italy did expert-endorsed debunking, compared to generic messaging, yield a modest uptick in the intent to vaccinate (+1.6%) and potentially bolster beliefs about protecting others.

Hence, an interpretation of Panizza's paper as a conceptual replication challenges the notion that communication explicitly backed by scientists or medical experts is a universal solution for increasing vaccination uptake. Here, "universal" refers to an approach that would

be effective under all conditions without exception. While the Italian findings indicate that expert-endorsed debunking can be beneficial, the UK results demonstrate that this outcome is not guaranteed. Rather than interpreting the UK data as a refutation of expert-endorsed debunking, the study underscores the need to investigate *when* and *how* such messages work, rather than assuming their global effectiveness. In relation to each other and interpreted against the background of recent work in medical anthropology, the findings from Ronzani, Panizza and their colleagues problematize two connected assumptions that underly information campaigns: first, that hesitancy stems primarily from misunderstanding scientific facts, and second, that it is a universal phenomenon unaffected by local context. The following sections sketches why and how these assumptions could be reconsidered.

## On the messengers: Perhaps there is not necessarily a knowledge deficit

For the research team, vaccination hesitancy served as a case study to analyze the role of expert endorsement in persuasion. They drew inspiration from theoretical frameworks such as the "elaboration likelihood model," which posits that the "credibility of the source, including perceived expertise, acts as a persuasive factor" (Panizza et al., 2023; Petty & Cacioppo, 1986). Vaccination hesitancy presents an interesting test case for this theory: It is often attributed to "misinformation," such as the "misplaced fear that [the vaccine] could cause autism in children (a case of false information)" (Panizza et al., 2023). This misinformation, according to the authors, could be "debunked" by "directly providing evidence from scientific studies demonstrating no correlation between vaccination and autism in children" (Panizza et al. 2023).

The main query for the authors was centered on accomplishing this most effectively. Their answer: by using a persuasive source. They reasoned that because experts and medical scientists are generally highly trusted, these groups would be the most effective in addressing the misunderstanding of science among vaccine-hesitant individuals. This position frames vaccine hesitancy as a conflict between science and ignorance—the former portrayed as unproblematic and the latter as flawed and in need of correction (Goldenberg, 2016). Hence, Panizza and his colleagues framed vaccine hesitancy in alignment with the "knowledge deficit model," which assumes that expert knowledge alone provides a sufficient foundation for resolving major public policy issues (Wynne, 1991). From this perspective, beliefs that contradict expert-endorsed directions should be corrected through education, science communication, or—indeed—"debunking."

However, historians, philosophers, and sociologists of science have long critiqued the assumptions underlying a knowledge deficit model (e.g., Hilgartner, 1990; Lewenstein, 1992; Miller, 2001; Wynne 1991; 1992). Publics whose beliefs or behaviors contradict scientific consensus are not necessarily rejecting scientific knowledge outright, these scholars point out. They are not "anti-science," but they mistrust science as an organization. The problem could therefore better be framed in terms of mistrust instead of ignorance. How does this apply to vaccine hesitancy? Philosopher Maya Goldenberg (2016) argues that "some of the previously secure relations of trust between science and the public that gave consensus statements their epistemic weight in the eyes of the lay public no longer hold" (p. 564). Based on ethnographic research, Goldenberg suggests that vaccine-hesitant parents do not unequivocally reject the scientific consensus on vaccines. Rather, they mistrust the priorities underpinning policy recommendations made by scientific experts. Regarding childhood immunization, Goldenberg notes that many parents she interviewed incorporated "established knowledge that immune responses do vary" and sought to address gaps in understanding causal or preceding events. These parents viewed the "presence of rare but serious adverse events as a safety priority rather than, as health officials see it, a reasonable risk" (Goldenberg, 2016, p. 566; p. 564). This reframes vaccine hesitancy not as the product of misinformation but as a reflection of diverging priorities: the safety of their children versus the safety of the population (Hobson-West, 2003; Hobson-West, 2007).

Goldenberg's argument cannot simply be applied wholesale to vaccine hesitancy during the COVID-19 pandemic, but it invites a reinterpre-

tation of the results of Panizza's experiment. Perhaps the findings reflect a lack of trust in the priorities of science rather than in consensus-based scientific knowledge itself. In the Italian study, participants were significantly more convinced that the vaccine protected others, yet there was no observable increase in vaccination uptake following the expert-endorsed messaging. This suggests that trust in scientific knowledge does not necessarily translate into trust in policy recommendations. That observation prompts an important question: Do the Italian and UK studies indicate that vaccine hesitancy stems from mistrust in the priorities of science rather than ignorance of its consensus? It would have been valuable to explore whether participants believed the vaccine was effective in preventing both disease occurrence and its transmission but still prioritized concerns about potential side effects in their individual cases. If this were true, it would suggest that people can simultaneously trust scientific knowledge while mistrusting public policies endorsed by scientists and medical experts.

It is unfair to characterize Panizza's experiment as purely one-directional. The researchers did account for participants' concerns when tailoring certain debunking messages. However, they also assumed that scientists held primary expertise on what mattered most when deciding to get vaccinated against COVID-19. This may be a mistake, as scholars in the medical humanities have suggested. Rather than focusing solely on which beliefs about vaccines underlie hesitation, these scholars argue for a more expansive approach to public engagement. Goldenberg, for instance, advocates moving toward a "dialogical" and "communicative" form of interaction that involves hesitant individuals in co-defining priorities (Goldenberg, 2016). Rather than presuming non-participation stems from ignorance to be corrected, dialogue could begin by recognizing local concerns and the communities whose health hesitant individuals prioritize. This form of public engagement fits recent calls for "community-based participatory research" in which stakeholders are involved equitably, and collaboratively articulate which research topics are of importance to specific communities with the aim of combining knowledge and action for social change to improve community health (Minkler & Chang, 2019; Wallerstein & Duran, 2010). From this perspective—and as Panizza and his colleagues themselves acknowledge—the specific context in which hesitant individuals live truly matters. The contrasting findings in Italy and the UK thus prompt a new research question: Does a single messaging strategy work across all contexts? We already know the answer, and it brings into sharp focus a second key assumption of the research group: That vaccination hesitancy is the same phenomenon everywhere.

## On the receivers: Perhaps vaccination hesitancy is not a global phenomenon

On the surface, it may seem plausible to view non-participants in vaccination campaigns as a single group: They all decline a readily available vaccine endorsed by governments and medical experts. Since vaccines and the diseases they address are global, one might assume vaccine hesitancy is similarly universal. From this perspective, it makes sense that Panizza and his colleagues expected their successful intervention in Italy to produce comparable outcomes in the United Kingdom. Yet in practice, it did not. In seeking an explanation, the researchers primarily scrutinized their quasi-experimental design and noted that factors such as the timing of vaccination campaigns and government communication efforts likely influenced the divergent results. Even more noteworthy is their suggestion that their intervention "might simply not be effective in certain populations." As they also remarked in their Italian study, they stress that in contrast to the UK, "the cultural characteristics of Italy make it peculiar in more than one way, and we may observe a certain degree of heterogeneity between populations" (Panizza et al. 2023). Hence, their findings imply that vaccination hesitancy is not uniformly distributed worldwide but rather shaped by distinct local contexts.

What does it mean, then, to consider vaccination hesitancy as a context-dependent phenomenon? Scholars in the medical humanities have long noted that hesitancy does not necessarily reflect direct opposition to government mandates or scientific consensus. Instead, it often arises from social interaction rather than purely rational deliberation. Anthropologist Elisa Sobo, for example, frames non-

participation as a form of solidarity within specific communities, calling it "refusal" rather than "resistance" to capture this nuance. Whereas "resistance" implies direct confrontation with institutional power, "refusal" operates as a generative act of solidarity—reaffirming a community's social fabric. In other words, communities that refuse vaccination may still acknowledge science but opt to uphold communal priorities or shared identity, rather than actively resisting external authority (Sobo, 2016).

This perspective invites for moving beyond portraying "vaccine refusers" as a monolithic group defined solely by shared (mis)beliefs about vaccines. If refusal represents an act of solidarity, the more pressing question becomes: Which community is the refuser reaffirming? Anthropologists such as Mia Hammerlin emphasize that non-participation often hinges on the "place" in which it occurs—a localized setting of proximal beliefs, customs, and daily interactions. In this sense, "local context" speaks to the historical, institutional, and cultural specificities that shape a community's way of living together, while "proximity" denotes not only physical closeness but also social interconnectedness and mutual dependence (Hammerlin, 2022). Rather than focusing solely on the reasons cited for not vaccinating, we might look more closely at *where*—and why there—groups of people opt out together. This helps situate vaccine refusal in its local context rather than in its abstract similarities across various publics, as Lawrence and her colleagues suggest (Lawrence et al., 2014).

Although focusing on the local "place" of vaccine refusal fosters a descriptive and actor-focused mode of analysis, it does not rule out the possibility of overarching mechanisms. Certain factors not yet included in existing taxonomies of vaccine uptake could explain why the same intervention succeeds in one setting but fails in another (MacDonald et al., 2022; Thomson et al., 2016). If such variables were identified in exploratory historical or ethnographic studies and then investigated across multiple locations, they might reveal one or more universal drivers of vaccine hesitancy—drivers that manifest differently depending on the "proximities" binding refusing communities. From this perspective, a one-size-fits-all framework could still work at a broad theoretical level, as long as it allows for the specific contextual factors that shape its practical effectiveness. This sensitivity facilitates a pragmatic and "situated" public health approach in response to vaccination refusal—one rooted in dialogue over priorities and an empirically traceable understanding of community values.

## Conclusion: Situated public health

This reflection articles proposes to interpret the conflicting outcomes in Italy and the United Kingdom as highlighting that the "knowledge deficit model" alone cannot explain why some communities remain hesitant regarding vaccination. If providing accurate, expert-endorsed information were universally sufficient, the researchers would very likely not find such a divergence in responses. Instead, their results underscore a "contextualist" perspective advocated by historians, sociologists, and philosophers of science: Neither "science" nor "the public" is a uniform entity. Both arise within specific historical and spatial environments, through interactions that co-produce expert knowledge and local understanding (Hilgartner,1990; Jasanoff, 2004; Lewenstein, 1992; Miller, 2001; Wynne 1991; 1992). In this light, as Maya Goldenberg argues, a *dialogical* approach to vaccine refusal becomes essential: Health organizations must recognize and genuinely engage with the priorities and concerns of those who refuse vaccination, rather than simply viewing them as uninformed (Goldenberg, 2016). Other medical humanities scholars, such as Heidi Lawrence and her colleagues, further emphasize that this dialogue ought to begin at the *local* level, with qualitative insight into populations' particular values and identities (Lawrence et al., 2014). This approach echoes community-based participatory research (CBPR), where researchers and local stakeholders collaborate at every stage—from defining the problem to designing interventions (Wallerstein & Duran, 2010). Such a contextualized approach to vaccine hesitancy could help align the diverse needs of individuals, communities, and nations in a way that top-down messaging alone cannot achieve.

Practically speaking, this is difficult. A situated response to vaccination refusal admittedly demands more work than rolling out expert-endorsed "debunking" materials. Yet,

if the scientific reasoning used to dispel vaccine doubts has failed to eliminate skepticism over the past two hundred years, "they are not going to start working now," as Lawrence and co-authors argue (Lawrence et al., 2014, p. 127). The findings from Panizza, Ronzani, and colleagues' research slightly refines this view: Although expert-backed messaging did not boost actual uptake in Italy, it significantly increased people's *intent* to vaccinate. Fixing the knowledge deficit may in fact be effective—sometimes, and in specific places. Yet, before using science-based misinformation campaigns as a panacea for vaccine refusal, it may be useful to first ask three questions: *Where* is hesitancy localized, *what* do the people in that setting have in common, and *how* could scientific reasoning align with their beliefs and practices?

Answers to these questions may clarify vaccine refusal but also pose a serious normative challenge for public health advocates: What if locally situated non-participation endangers overall population health? Measures like restricting daycare access, as public health ethicists such as Pierik and Verweij (2024) propose under John Stuart Mill's no-harm principle, might be warranted when voluntary uptake fails. However, these interventions risk undermining the communal ties that refusal often reinforces. Consequently, health officials must ensure that local solidarities are not dismissed, while still prioritizing broader public health goals. A contextualized understanding of refusal could help them navigate this tension when co-designing interventions to maximize participation in vaccine campaigns. Perhaps community leaders, rather than distant experts, are the most convincing messengers; trusted clerics could very well be more persuasive than health officials. Maybe religious doubts should be met with faith-based reassurance. And if trust in governments is low, civil society organizations might have better luck administering vaccines. These are mere suggestions in need of empirical verification.

In the end, I would say that the main oversight in Panizza's approach was assuming that the arguments driving hesitancy in Italy, identified through a pre-trial explorative study, would directly apply to the UK. I wonder whether a more context-sensitive approach—attuned to the diversity of cultural, social, and historical specifics, or at least informed by responses from UK participants—might have produced different results. Such adjustments could have made their paper as a conceptual replication even more compelling. Nevertheless, if we accept the possibility that vaccine hesitancy is rooted in specific local realities, then grouping participants by country borders is likely not the most meaningful analytic strategy. The real takeaway from these studies is that effective public health interventions should probably do more than correct a supposed misunderstanding of science or amplify expert authority. They should also resonate with the decisions hesitant individuals make—or don't make—within the places they live. Ultimately, Panizza and his colleagues' null result is not a "failed experiment". It is a gesture in the direction of situated public health interventions.

## Acknowledgements

I have greatly benefited from interacting with Elena Conis and the fellows from the Kavli Center for Ethics, Science, and the Public at the University of California at Berkeley. I would also like to thank David Jones and Allan Brandt from Harvard History of Science; Hans van Vliet from the Dutch Institute for Public Health and the Environment; Noortje Jacobs, Timo Bolt, and Laura Hartman from the Erasmus Medical Center in Rotterdam; and Ralf Futselaar from the Erasmus University in Rotterdam for discussing my interpretation of vaccine refusal at various occasions. The three challenging reviews from the *Journal of Trial and Error* have been invaluable to sharpen and nuance the final text. Yet above all, I would like to thank Folco Panizza and his colleagues for ensuring that their findings did not end up in a file drawer.

## References

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99. https://doi.org/10.1016/j.jesp.2015.10.002

European Commission. (2024). *Policy, expertise, and trust in action | PERITIA Project | Results | H2020.* CORDIS | European Commis-

sion. https://cordis.europa.eu/project/id/870883/results

Flatt, A., Vivancos, R., French, N., Quinn, S., Ashton, M., Decraene, V., Hungerford, D., & Taylor-Robinson, D. (2024). Inequalities in uptake of childhood vaccination in England, 2019-23: Longitudinal study. *British Medical Journal, 387*, Article e079550. https://doi.org/10.1136/bmj-2024-079550

Goldenberg, M. J. (2016). Public misunderstanding of science? Reframing the problem of vaccine hesitancy. *Perspectives on Science*, *24*(5), 552–581. https:/ / doi.org/ 10.1162/POSC_a_00223

Hammerlin, M.-M. (2022). This is home: Vaccination hesitancy and the meaning of place. *Ethnologia Scandinavica, 52*, 202-220.

Hilgartner, S. (1990). The dominant view of popularization: Conceptual problems, political uses. *Social Studies of Science*, *20*(3), 519–539. http://doi.org/10.1177/030631290020003006

Hobson-West, P. (2003). Understanding vaccination resistance: Moving beyond risk. *Health, Risk & Society, 5*(3), 273–283. https://doi.org/10.1080/1369857031000160697 8

Hobson-West, P. (2007). 'Trusting blindly can be the biggest risk of all': Organised resistance to childhood vaccination in the UK. *Sociology of Health & Illness, 29*(2), 198–215. https://doi.org/10.1111/j.1467-9566.2007.00544.x

Jasanoff, S. (Ed.). (2004). *States of knowledge: The co-production of science and social order*. International Library of Sociology. Routledge.

Lawrence, H. Y., Hausman, B. L., & Dannenberg, C. J. (2014). Reframing medicine's publics: The local as a public of vaccine refusal. *Journal of Medical Humanities*, *35*(2), 111–129. https://doi.org/10.1007/s10912-014-9278-4

Lewenstein, B. (Ed.) (1992). *When science meets the public: Proceedings of a workshop organized by the American Association for the Advancement of Science, Committee on Public Understanding of Science and Technology, February 17, 1991, Washington, DC*. American Association for the Advancement of Science. https://ecommons.cornell.edu/bitstream/handle/1813/70154/Lewenstein.1992.When%20Science%20Meets%20Public.pdf?sequence=3

Miller, S. (2001). Public understanding of science at the crossroads. *Public Understanding of Science, 10*(1), 115–120. https://doi.org/10.3109/a036859

Minkler, M., & Chang, C. (2019). Community-Based Participatory Research: A promising approach for studying and addressing immigrant health." In Marc B. Schenker, Xóchitl Castañeda, & Alfonso Rodriguez-Lainz (Eds.), *Migration and Health* (pp. 361–376). University of California Press. https://doi.org/10.1525/9780520958494-020.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Panizza, F., Ronzani, P., Martini, C., Savadori, L., & Motterlini, M. (2023). Medical expert endorsement fails to reduce vaccine hesitancy in U.K. residents. *Journal of Trial and Error*. https://doi.org/10.36850/e15

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In Richard E. Petty & John T. Cacioppo (Eds.), *Communication and persuasion: Central and peripheral routes to attitude change* (pp. 1–24). Springer. https://doi.org/10.1007/978-1-4612-4964-1_1

Pierik, R. & Verweij, M. (2024). *Inducing immunity? Justifying immunization policies in times of vaccine hesitancy*. MIT Press. https://books.google.com/books?hl=nl&lr=&id=nCfHEAAAQBAJ&oi=fnd&pg=PR5&dq=inducing+immunity+verweij&ots=NYjzF6JNyW&sig=c9BeX659N8NPL5OSOeU8bWStbnw

Porter, D. (2005). *Health, civilization and the state: A history of public health from ancient to modern times*. Routledge.

Ronzani, P., Panizza, F., Martini, C., Savadori, L., & Motterlini, M. (2022). Countering vaccine hesitancy through medical expert endorsement. *Vaccine, 40*(32), 4635–4643. https://doi.org/10.1016/j.vaccine.2022.06.031

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100. https://doi.org/10.1037/a0015108

Seither, R. (2024). Coverage with selected vaccines and exemption rates among children in kindergarten — United States, 2023–24 School Year. *Morbidity and Mortality Weekly Report*, *73*(41), 925-932. https://doi.org/10.15585/mmwr.mm7341a3

Sobo, E. J. (2016). Theorizing (vaccine) refusal: Through the looking glass. *Cultural Anthropology, 31*(3), 342–350. https://doi.org/10.14506/ca31.3.04

van Lier, E. A., Hament, J.-M., Knijff, M., Westra, M., & Giesbers, H. (2024). Vaccinatiegraad Rijksvaccinatieprogramma Nederland, verslagjaar 2024. Rijksinstituut voor Volksgezondheid en Milieu (RIVM). https:/ / doi.org/ 10.21945/RIVM-2024-0044

Vargha, D., & Wilkins, I. (2023). Vaccination and pandemics. *Isis, 114*(S1), S50–S70. https://doi.org/10.1086/726980

Wallerstein, N., & Duran, B. (2010). Community-Based Participatory Research contributions to intervention research: The intersection of science and practice to improve health equity. *American Journal of Public Health, 100*(S1), S40–46. https://doi.org/10.2105/AJPH.2009.184036

Wynne, B. (1991). Knowledges in context. *Science, Technology, & Human Values, 16*(1), 111–121. https:/ / doi.org/ 10.1177/ 016224399101600108

Wynne, B. (1992). Misunderstood misunderstanding: Social identities and public uptake of science. *Public Understanding of Science, 1*(3), 281–304. https:/ / doi.org/ 10.1088/ 0963-6625/1/3/004

**Correction notice**
Incorrect Special Issue Labeling (Article erroneously excluded): This article was previously not labeled as part of a special issue due to an error. This has now been corrected.

# Digital Nudges: A Reflection on Challenges and Improvements Inspired by the Gloria Adherence Subproject

David Grüning [1, 2]

Rapid technological development allows for ever new opportunities to nudge individuals' behavior and knowledge digitally. The Gloria Adherence Subproject by Horne et al. (2022) implements such a digital nudge via a mobile device aiming at medication adherence. Besides methodological and practical shortcomings outlined by the authors themselves, the adherence nudge used might have had conceptual weaknesses. In the present article, I reflect on three prominent challenges of digital nudges in general and an emerging redemption of the nudging-concept in the form of so-called boosts. Both reflections inform the evaluation of the outcomes of the Gloria Adherence Subproject and suggest specific actions for optimization for future project retrials or conceptual replications by other scientists.

**Keywords** *nudge, digital, types, challenges, boost*

[1] Heidelberg University & GESIS - Leibniz Institute for the Social Sciences
[2] Center of Trial & Error, Utrecht, the Netherlands

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
January 23, 2023
**Accepted**
March 21st, 2023
**Published**
April 6th, 2023

**Correspondence**
Heidelberg University & GESIS - Leibniz Institute for the Social Sciences
gruening@trialanderror.org



With rapid technological advancements researchers and practitioners are offered new and improved opportunities to intervene with individuals' behavior and knowledge digitally. Whole industry sectors focus on digital interventions that aim to increase the flourishing and well-being of the receiver (e.g., apps like *one sec*[1] or *Structured*[2]). Psychological research has also focused on developing and scrutinizing digital nudges to advance a multitude of fields concerned with but not limited to prosociality offline and online (e.g. Matias, 2019; Tyler et al., 2021; for a conceptual review, Grüning et al., 2022), misinformation (e.g. Donovan & Rapp, 2020; Guess et al., 2020; Lutzke et al., 2019; for a review, Kozyreva et al., 2022), consumption of digital contents (e.g. Grüning, Riedel, & Lorenz-Spreen, 2023), and medical health behavior (e.g. Glasgow et al., 2021; Horne et al., 2022; Luong et al., 2021).

In a similar vein as the last example, a digitally directed nudge was tested by Hartman et al. (2021) in the *Gloria Adherence Subproject* to increase medication adherence in the form of the daily pill intake of patients with a chronic inflammatory disease (e.g., rheumatoid arthritis). As the authors already outline methodological

[1] one-sec.app/, an app to decrease automatic digital consumption and promote deliberate consumption.
[2] structured.app/, an app to promote structured day-planning.

## Companion Article

Hartman, L., Kok, M. R., Molenaar, E., Griep, N., Van Laar, J. M., Van Woerkom, J. M., F., A. C., Raterman, H. G., Ruiterman, Y. P. M., Voshaar, M. J. H., Redol, J., Pinto, R. M. A., Klausch, L. T., Lems, W. F., & Boers, M. (2021). The Gloria Adherence Subproject: Problems and randomization mistakes. *Journal of Trial and Error, 2*(1), 50–55. https://doi.org/10.36850/e6

and practical issues of the study in detail, in the present reflection article, I will focus on a higher-level scope of discussion. That is, I aim for a broader reflection on nudges that are linked to technology as their medium, namely, digital devices or online tools. I will (1) outline two different types of digital nudges, (2) review challenges of these unsupervised nudges, and (3) suggest an improvement of them that is already researched and applied in different digital spaces: boosting (e.g. Hertwig, 2017; Hertwig & Grüne-Yanoff, 2017; Lorenz-Spreen et al., 2021; Reijula & Hertwig, 2022). In sum, the present paper is meant as a reflection on digital nudges per se, stimulated by the well-reflected use of this kind of intervention in the Gloria Adherence Subproject.

## Nudging intervention in the Gloria Adherence Subproject

Hartman et al. (2021) in the Gloria Adherence Subproject designed a between-subjects study with one experimental and one control group to test the effect of an adherence nudge on medication intake in patients with chronic inflammatory diseases. In the former group, patients were equipped with an app on their mobile device that reminded them daily to take their required medication, in the form of a pill, if they had not already done so. The control group was presented with no such intervention. Pill intake was measured via screening the removal of the electronic cap of the used pill bottles. The authors found no effects of the adherence nudge on patients' consistency in taking their medication. While Hartman and colleagues already consider practical and methodological shortcomings of the subproject in exemplary extensiveness, I want to address potential problems of the intervention itself and of digital nudges more generally. That is, the null results of the adherence nudging demonstrated in the Gloria Adherence Subproject stimulate further, more general, questions about the limits of digital nudging which are the focus of the present paper. I reflect on three challenges of digital nudging prominently discussed in intervention research, and on an emerging and commendable improvement of the nudging-concept named boosting. Beforehand, however, a general outline of digital nudges is needed.

## Two types of digital nudges and information environments

Many researchers (Bail, 2021; Thomas, 1983; Gigerenzer et al., 2011; Gigerenzer & Selten, 2002; Simon, 1956) already pointed out that an individual must learn about an environment's properties in order to function within it effectively. This realization is also important for understanding the (in)effectiveness of nudging interventions under certain circumstances. Environments vary between two extremes: *kind* and *wicked*. Kind environments provide feedback that is mostly accurate and direct. Wicked environments are based upon opaque variables and their moderators, which are interconnected and subject to changes across time. As a result, any feedback in wicked environments is volatile, confounded with moderators, and commonly presented with time delays (Hogarth, 2001). Kind environments – like board games and most sports –lead to a positive learning curve for the individual through their reliable feedback utilization. Even relatively uninformed decision-makers can easily adapt to kind environments by trial and error behavior generating instructive feedback. By repeatedly taking action, people can reliably learn about the environment's conditions and manage their intentions and actions accordingly. In contrast, wicked environments – like weather and climate, the stock market, or social media – can make people spiral into grave misconceptions about their environment. The result are misguided judgments and decisions (e.g. Denrell & March, 2001; Einhorn & Hogarth, 1978; Feiler et al., 2012; Koehler & Mercer, 2009).

Like environments, Grüning, Panizza, and Lorenz-Spreen (2023) propose that digital nudges can be categorized into two distinct types. The first kind are *behavioral nudges* instantiating behavior change. Research on this nudging type is abundant and its tested environments are ample (see e.g. Daley et al., 2018; Glanz et al., 2017; Pennycook et al., 2021; Voelkel et al., 2021). Behavioral nudges aim at *pushing* individuals towards a specific action. For example, in the social media context, based on insights by Pennycook et al. (2021), Twitter warns about sharing a link without having accessed it, in order to reduce the distribution of shared misinformation. The second

type concerns *informative nudges*. These interventions provide people with information about their environment, thereby increasing the transparency about decisive environmental variables and processes. Two examples of this intervention type are illustrative of its characteristics. First, Guess et al. (2020) developed an intervention to promote users' digital media literacy. The authors presented social media users with tips about finding additional online information to double-check news and presumable facts circulated on social media (see a review of the intervention, Prosocial Design Prosocial Design Network, n.d.). Second, every time a user attempts to open an application on their mobile phone, the app *one sec* (one-sec.app) can inform them about their overall number of daily consumption attempts, enhancing transparency of their digital behavior. The medication adherence nudge implemented by Hartman et al. (2021) in the Gloria Adherence Subproject can be classified as a behavioral intervention.

## Challenges of digital nudges

As rich and broad as the empirical evidence for digital nudging is, this type of intervention for behavioral change and information gain is confronted with core challenges to its general concept. I outline three prominent problems of the nudging-concept in the following sections.

### Undermining agency

Nudges, especially for behavior change, are meant as automatic or subtle enhancements of certain actions and behavioral patterns. To have the anticipated effect, such interventions have to be cast upon a user unknowingly and unwillingly. Nudges do not allow the addressee to have a say, that is, to choose which nudges to use or not to use for themselves, substantially undermining users' agency (for an extensive discussion Hertwig & Ryall, 2020). This is further problematized because nudges do not work through transparency. The opposite is the case: nudges commonly require a certain degree of intransparency toward the user to be successful in affecting their behavior. Addressees of the typical nudge are not aware that they are nudged nor are they able to influence the nudging to their benefit. In short, nudges run the grave risk of patronizing their addressees and are, on top of this, (mostly) undetectable and uncontrollable for them, denying any chance for users to become aware of their passivity. In the Gloria Adherence Subproject specifically, the message nudge was predetermined in presentation (i.e., when displayed) and content (i.e., nudging message) format, allowing no possibility for participants to adapt it to their individual needs (e.g., their daily routine).

### Effect decay by time

While nudges are resource-efficient and simple to implement, their effects decay relatively quickly, as shown in diverse fields like cognitive biases (Grüning et al., 2022), judgement accuracy (Lorenz-Spreen et al., 2021) and memory (Trammell & Valdes, 1992). Nudges need consistent reinforcement. For an illustrative example, Grüning, Panizza, and Lorenz-Spreen (2023) tested the nudging effect of *one sec,* an app to reduce users' digital consumption. Every time a user attempted to open another app on their mobile device (e.g., to browse Twitter or TikTok) *one sec* nudged the user to dismiss this app opening by reminding them of what they were about to do. This nudge, the authors demonstrate, is immensely effective. However, its effectiveness relies on the repeated and constant nudging of its users. Applying this insight to the Gloria Adherence Subproject, it is notable that the nudge of pill-intake by adherence message only occurred once per day, namely if the medication had not been used after a certain time point. Due to their simple and straightforward design, nudges are especially useful for resource-efficient short term effects. However, this also means that their impact runs out rather quickly and is susceptible to moments of inattention. The Gloria Adherence Project may have suffered from this nudge characteristic by providing their nudge too scarcely.

### Risks in wicked environments

Nudges are especially useful in environments that provide a straightforward and unmasked feedback loop, that is, kind environments. Here nudging can support individuals to in-

creasingly show behavior that promotes positive effects for them (e.g., in well-being, income, and personal relationships). However, applying simple nudges in more complex environments with influential and interconnected moderators (that is, wicked environments), can not only prove ineffective but can also backfire for the individual. For instance, the accuracy nudge against digital misinformation that was recently suggested (e.g. Pennycook et al., 2021; Lorenz-Spreen et al., 2021) simply prompts users to be aware of cues which are characteristic of misinformation, like a dubious information source. However, the accuracy nudge may not help or may even do harm if the indicators that people use to judge the accuracy of a piece of information are non-diagnostic. In this example, sources can be intentionally constructed to increase their perceived credibility and distribute misinformation more effectively. In wicked and malignant environments, blindly following simple nudges can lead to grave misinterpretations and result in maladaptive behaviors. For the Gloria Adherence Project, participants are also nudged in the context of a complex environment, in which diverse moderators (e.g., interpersonal events, fluctuations in health, and individually different routines) can interfere with the nudge's effectiveness.

While nudging has a long scientific tradition and has accumulated a body of evidence promoting its effectiveness, the intervention's challenges restrict the viability and, most importantly, applicability of digital nudges.

## Redemptions by boosting

Confronted with such challenges of common nudging interventions, we invite the wider scientific community to consider a more granular inquiry into, and application of, interventions for behavioral change and information gain. One prominent alternative to nudging, or rather an advancement of the nudging-concept, is *boosting* (e.g. Hertwig, 2017; Hertwig & Grüne-Yanoff, 2017). Instead of merely highlighting a preferred action or behavioral pattern (e.g., taking the daily pill), boosts foster individuals' knowledge of and competencies for the respective positive behavior. As a result, boosting interventions allow effective self-directed action (e.g. Lorenz-Spreen et al., 2021). A successful digital intervention moves beyond mere nudging and, additionally, fosters the user's ability to navigate the environment themselves.

A meaningful illustration of successful boosting are inoculation strategies used to address online manipulation (e.g., polarization or product targeting). For example, Lorenz-Spreen et al. (2021) aimed at improving users' competence to detect manipulative strategies online by not just priming in users' minds that individuals are susceptible to manipulative online content but also allowing participants to reflect about their individual susceptibility to such manipulation. Specifically, the authors presented respondents with different manipulative advertisements targeting extraverted or introverted consumers. Supporting participants to reflect about this personality trait in relation to themselves increased their accuracy in detecting the presented advertisements' targeting strategies compared to just priming the targeted personality trait in the individuals' mind.[3] Many more illustrations of inoculation (e.g., as digital games Basol et al., 2020; Harrop et al., n.d.) and of other forms of boosting – for example media literacy (Guess et al., 2020) and rebutting science denialism (Schmid & Betsch, 2019) – exist. Research also attests to boosting's effects not only in the short-term, but also long term (e.g. Maertens et al., 2021), compared to nudging.

In the following, I revisit the three previously-mentioned challenges of nudging, as applied to the Gloria Adherence Subproject and offer remedies by boosting application. For a potential future retrial of the Gloria Adherence Subproject, I suggest not just scrutinizing the methodological and practical circumstances as described by the authors but also integrating a more effective, long-term oriented version of their adherence intervention. Instead of merely reminding patients to take their daily medication, researchers could provide them with additional information. For instance, the mobile device could provide an accessible information glossary of the pill, its effects, and

[3] Note that this study only concerned short term effects on detecting manipulation strategies. According to the boosting concept, we should also observe long term effects as reflecting about one's personal susceptibility to manipulation should lead to strengthening a user's competence sustainably. In comparison, the mere nudging effect should be quickly exhausted.

the importance of taking it at repeated and consistent intervals. Further, patients could be informed about their past pill intake behavior, being motivated by streaks of adherence (i.e., *n* days of constant and on-time medication adherence) and sensitized by failures to sustain the streak. Allowing patients to understand the reasons for and consequences of the nudged behavior and offering them the opportunity to screen their behavioral performance should foster patients' competencies, with the long-term effect that pill intake (and any other medication adherence for that matter) grows more reliable. Via the above-described suggestion of a boosting advancement, participants would be more aware of how the medication intervention works and would be allowed to adapt the intake (in a limited way) to their individual routine, enabling them to have more control over the intervention by themselves. Further, advancing the adherence nudge should also address issues of time decay. Informing participants effectively about the idea of the implemented adherence intervention should raise their awareness of the importance of medication intake and also boost the cognitive prominence of having to take the medication. Lastly, fostering the participants' own competencies regarding their medication intake (e.g., being more informed and in more control) should allow them to adapt more dynamically to issues that arise unexpectedly. This increase in effective navigation of one's environment should also help individuals to avoid backlash effects – for example, from unexpected personal events interfering with the planned medication intake. In summary, boosting has the potential to improve participants' agency in the experimental group, intercept effect decay over time, and make the intervention more robust against complex moderators in the environment.

## Conclusion

In the present reflection article, I selectively revisited the scientific landscape on digital nudges, their formats, challenges, and a commendable improvement of its concept. Digital nudges can be distinguished into aiming at behavioral change or at gaining information. Their most prominent challenges are ethical limitations (in terms of individual agency), their effect decay over time, and ineffectiveness or even backlash in more complex environments. Boosting as an advancement of the nudging-concept is robust against all three of these problems. In case of a planned retrial or conceptual replication of the Gloria Adherence Subproject, Hartman et al. (2021) and other scientists could advance their medication adherence intervention from a mere nudging to a boosting format, thereby increasing the probability of positive medication adherence effects.

## References

Bail, C. (2021). *Breaking the social media prism.* Princeton University Press. (See p. 2).

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. Journal of Cognition, 3(1), 2. https:/ / doi.org/ 10.5334/ joc.91 (see p. 4).

Daley, M. F., Narwaney, K. J., Shoup, J. A., Wagner, N. M., & Glanz, J. M. (2018). Addressing parents' vaccine concerns: A randomized trial of a social media intervention. American Journal of Preventive Medicine, 55(1), 44–54. https://doi.org/10.1016/j.amepre.2018.04.010 (see p. 2).

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. Organization Science, 12(5), 523–538. https://doi.org/10.1287/orsc.12.5.523.10092 (see p. 2).

Donovan, A. M., & Rapp, D. N. (2020). Look it up: Online search reduces the problematic effects of exposures to inaccuracies. Memory & Cognition, 48(7), 1128–1145 (see p. 1).

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. Psychological Review, 85(5), 395–416. https:/ / doi.org/ 10.1037/ 0033-295X.85.5.395 (see p. 2).

Feiler, D. C., Tong, J. D., & Larrick, R. P. (2012). Biased judgment in censored environments. Management Science, 59(3), 573–591. https://doi.org/10.1287/mnsc.1120.1612 (see p. 2).

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). Heuristics: The foundations of adaptive behavior. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199744282.001.0001 (see p. 2).

Gigerenzer, G., & Selten, R. (Eds.). (2002).

Bounded rationality: The adaptive toolbox. MIT Press. (See p. 2).

Glanz, J. M., Wagner, N. M., Narwaney, K. J., Kraus, C. R., Shoup, J. A., Xu, S., O'Leary, S. T., Omer, S. B., Gleason, K. S., & Daley, M. F. (2017). Webbased social media intervention to increase vaccine acceptance: A randomized controlled trial. Pediatrics, 140(6), 1–9. https://doi.org/10.1542/peds.2017-1117 (see p. 2).

Glasgow, R. E., Knoepke, C. E., Magid, D., Grunwald, G. K., Glorioso, T. J., Waughtal, J., Marrs, J. C., Bull, S., & Ho, P. M. (2021). The NUDGE trial pragmatic trial to enhance cardiovascular medication adherence: Study protocol for a randomized controlled trial. Trials, 22(1), 1–16 (see p. 1).

Grüning, D. J., Mata, A. O. P., & Fiedler, K. (2022). First exploration of the prediction-comprehension bias. https:/ / doi.org/ 10.31234/osf.io/qp894 (see pp. 1, 3).

Grüning, D. J., Panizza, F., & Lorenz-Spreen, P. (2023). The importance of informative interventions in a wicked environment. The American Journal of Psychology, 135(5), 439–442. https://doi.org/10.5406/19398298.135.4.12 (see pp. 2, 3).

Grüning, D. J., Riedel, F., & Lorenz-Spreen, P. (2023). Directing smart phone use through the selfnudge app one sec. Proceedings of the National Academy of Sciences, 120(8), 2213114120. https:/ / doi.org/ 10.1073/ pnas.2213114120 (see p. 1).

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. Proceedings of the National Academy of Sciences, 117(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117 (see pp. 1, 3, 4).

Harrop, I., Roozenbeek, J., Madsen, J., & van der Linden, S. (n.d.). Inoculation can reduce the perceived reliability of polarizing social media content. International Journal of Communication (see p. 4).

Hartman, L., Kok, M. R., Molenaar, E., Griep, N., Van Laar, J. M., Van Woerkom, J. M., F., A. C., Raterman, H. G., Ruiterman, Y. P. M., Voshaar, M. J. H., Redol, J., Pinto, R. M. A., Klausch, L. T., Lems, W. F., & Boers, M. (2021). The Gloria Adherence Subproject: Problems and randomization mistakes. Journal of Trial and Error, 2(1), 50–55. https://doi.org/10.36850/e6 (see pp. 1, 2, 3, 5).

Hertwig, R. (2017). When to consider boosting: Some rules for policy-makers. Behavioural Public Policy, 1(2), 143–161. https://doi.org/10.1017/bpp.2016.14 (see pp. 2, 4).

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. Perspectives on Psychological Science, 12(6), 973–986. https://doi.org/10.1177/1745691617702496 (see pp. 2, 4).

Hertwig, R., & Ryall, M. D. (2020). Nudge versus boost: Agency dynamics under libertarian paternalism. The Economic Journal, 130(629), 1384–1415. https://doi.org/10.1093/ej/uez054 (see p. 3).

Hogarth, R. M. (2001). Educating intuition. Chicago University Press. https://press.uchicago.edu/ucp/books/book/chicago/E/bo3624460.html (see p. 2).

Horne, B. D., Muhlestein, J. B., Lappé, D. L., May, H. T., Le, V. T., Bair, T. L., Babcock, D., Bride, D., Knowlton, K. U., & Anderson, J. L. (2022). Behavioral nudges as patient decision support for medication adherence: The encourage randomized controlled trial. American Heart Journal, 244, 125–134. https://doi.org/10.1016/j.ahj.2021.11.001 (see p. 1).

Koehler, J. J., & Mercer, M. (2009). Selection neglect in mutual fund advertisements. Management Science, 55(7), 1107–1121. https:/ / doi.org/ 10.1287/ mnsc.1090.1013 (see p. 2).

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., & Hertwig, R. (2022). Toolbox of interventions against online misinformation and manipulation. https://doi.org/10.31234/osf.io/x8ejt (see p. 1).

Lorenz-Spreen, P., Geers, M., Pachur, T., Hertwig, R., Lewandowsky, S., & Herzog, S. (2021). Boosting people's ability to detect microtargeted advertising. Scientific Reports, 11(1), 15541. https:/ / doi.org/ 10.1038/ s41598-021-94796-z (see pp. 2, 3, 4).

Luong, P., Glorioso, T. J., Grunwald, G. K., Peterson, P., Allen, L. A., Khanna, A., Waughtal, J., Sandy, L., Ho, P. M., & Bull, S. (2021). Text message medication adherence reminders automated and delivered at scale across two institutions: Testing the nudge system: Pilot study. Circulation: Cardiovascular Quality and Outcomes, 14(5), 007015. https://doi.org/10.1161/CIRCOUTCOMES.120.007015 (see p. 1).

Lutzke, L., Drummond, C., Slovic, P., & Árvai, J. (2019). Priming critical thinking: Simple interventions limit the influence of fake news about climate change on facebook. Global Environmental Change, 58, 101964. https://doi.org/10.1016/j.gloenvcha.2019.101964 (see p. 1).

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. Journal of Experimental Psychology: Applied, 27(1), 1–16. https://doi.org/10.1037/xap0000315 (see p. 4).

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. Proceedings of the National Academy of Sciences, 116(20), 9785–9789. https://doi.org/10.1073/pnas.1813486116 (see p. 1).

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. Nature, 592(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2 (see pp. 2, 4).

Prosocial Design Network. (n.d.). List tips for checking accuracy of shared headlines: Reduce the spread of mis- and disinformation. https://www.prosocialdesign.org/library/list-tips-for-checking-accuracy-of-shared-headlines (see p. 3).

Reijula, S., & Hertwig, R. (2022). Self-nudging and the citizen choice architect. Behavioural Public Policy, 6(1), 119–149. https://doi.org/10.1017/bpp.2020.5 (see p. 2).

Schmid, P., & Betsch, C. (2019). Effective strategies for rebutting science denialism in public discussions. Nature Human Behaviour, 3(9), 931–939. https://doi.org/10.1038/s41562-019-0632-4 (see p. 4).

Simon, H. A. (1956). Rational choice and the structure of the environment. Psychological Review, 63(2), 129–138. https://doi.org/10.1037/h0042769 (see p. 2).

Thomas, L. (1983). The youngest science: Notes of a medicine watcher. Viking. (See p. 2).

Trammell, N. W., & Valdes, L. A. (1992). Persistence of negative priming: Steady state or decay? Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(3), 565–576 (see p. 3).

Tyler, T., Katsaros, M., Meares, T., & Venkatesh, S. (2021). Social media governance: Can social media companies motivate voluntary rule following behavior among their users? Journal of Experimental Criminology, 17(1), 109–127. https://doi.org/10.1007/s11292-019-09392-z (see p. 1).

Voelkel, J. G., Chu, J., Stagnaro, M., Mernyk, J. S., Redekopp, C., Pink, S. L., Druckman, J. N., Rand, D. G., & Willer, R. (2021). Interventions reducing affective polarization do not necessarily improve antidemocratic attitudes. Nature, 7, 55–64. https://doi.org/10.1038/s41562-022-01466-9 (see p. 2).

**Correction notice**
Incorrect Special Issue Labeling (Article erroneously excluded): This article was previously not labeled as part of a special issue due to an error. This has now been corrected.

# Cognitive Functions, Mood and Sleep Quality after Two Months of Intermittent Fasting

Maja Batorek[1], Ivana Hromatko[2]

Intermittent fasting is being popularized as a method beneficial not only for weight loss, but also for overall psychological functioning and well-being. However, there is only a handful of studies examining the latter claims. The aim of this open-label study was to contribute to the understanding of the relationship between fasting-based diets, and cognitive functions and other mental health factors such as mood and sleep quality. The research was conducted on a sample of 105 healthy volunteers who were placed in either the experimental (fasting) group ($n$ = 76) or the control (no change in diet regimen) group ($n$ = 29). For a period of 2 months, the experimental group adhered to a time-restricted eating (TRE) form of intermittent fasting: Participants were instructed to fast from eating or drinking for 16 hours per day. Participants in the control group did not adhere to any specific dietary regimen. Cognitive functioning (attention, memory, working memory and executive functions), as well as sleep quality and several mood dimensions (anxiety, depression, fatigue, hostility, friendliness, cheerfulness, concentration, energy) were measured across three time points: Prior to the beginning of the study, and one month and two months later, respectively. Results showed no significant group x time point interactions on any of the measures. In conclusion, the results of this study do not corroborate the notion that TRE regimen significantly influences cognitive functions, mood or sleep of healthy individuals. While fasting-based diets successfully regulate weight, the claims regarding their beneficial effect on psychological functioning in non-clinical populations are yet to be proven.

**Keywords** *intermittent fasting, cognitive functions, mood, sleep quality, mental health*

[1]University of Zagreb, Zagreb, Croatia

[2]Faculty of Humanities and Social Sciences, Dept. of Psychology, University of Zagreb, Zagreb, Croatia

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
July 4, 2023
**Accepted**
April 11, 2024
**Published**
August 11, 2024

**Correspondence**
University of Zagreb, Faculty of Humanities and Social Sciences, Dept. of Psychology
ihromatko@ffzg.hr



Intermittent fasting is an umbrella term for several related dietary regimens that are based on the cyclical alternation of periods of usual food intake and periods of fasting, that is, periods in which food intake is significantly reduced or completely eliminated (Mattson et al., 2017). In contrast to caloric restriction dietary regimens, the emphasis in fasting is not on reducing caloric intake, but on limiting the period of feeding (Mattson & Longo, 2014; Khedkar, 2020). Three fasting regimens are most common. These are: whole day fasting, that is, complete abstinence from food 1 or 2 days a week; alternate day fasting, that is, limiting food consumption to a maximum of 25% of usual intake every other day; and time restricted eating (TRE) in which the change of fasting and feeding periods takes place within 1 day (Patterson & Sears, 2017). The main premise underlying weight loss with intermittent fasting is that a person cannot fully compensate for the energy deficit they achieve during the fasting period (Rynders et al., 2019). Reviews (Patterson et al., 2015; Welton et al., 2020) show that various fasting regimens are indeed effective

**Companion Article**

Schleim (2024)

*Cognitive or Emotional Improvement through Intermittent Fasting? Reflections on Hype and Reality*

DOI: 10.36850/4032-1db2

## Take-home message

The claims regarding the beneficial effects of intermittent fasting on various aspects of physical and psychological functioning might have been exaggerated. The evidence regarding its effects on mental health and cognitive functions is scarce, and this open-label study showed no changes in attention, memory, working memory, executive functions, mood or sleep quality after two months of intermittent fasting. Since no dietary regime is without potential adverse reactions, we call for caution in disseminating unsubstantiated claims regarding the benefits of intermittent fasting which transcend those related to weight loss.

method of losing weight, as statistically significant weight loss was found in a majority of the studies. TRE is currently the most popular form of intermittent fasting because the fasting period lasts much shorter than those in the other forms (Moro et al., 2016).

It has been shown that intermittent fasting interventions also lead to improvements in symptoms of metabolic syndrome such as insulin resistance, hypertension, and inflammation (Harvie et al., 2010; Moro et al., 2016). However, intermittent fasting drew attention among scientists in fields outside of nutrition, due to reports of its positive effects on psychological well-being (Patterson & Sears, 2017). Early studies have shown that caloric restriction in rodents leads to improved maze learning (Idrobo et al., 1987) and has a protective effect against cognitive decline in old age (Ingram et al., 1987). Relatively recent research shows that restricting the feeding period to 8 hours promotes neurogenesis and has protective effects after a stroke (Manzanero et al., 2014), and it also affects neurotrophin regulation (Marosi & Mattson, 2014). The results of a longitudinal study demonstrated that a group that adhered to caloric restriction for 2 years achieved greater improvements in working memory test as compared to the control group (Leclerc et al., 2020). People who consistently adhered to intermittent fasting performed better on cognitive tests than those who adhered poorly or not at all (Ooi et al., 2020). At the same time, a recent systematic review (Benau et al., 2021) showed that short-term fasting (3-48 hours) is more likely to be associated with impairment rather than improvement in higher-order cognitive functions.

As for the affective domain, excessive food intake has been associated with an increased risk of stroke and neurodegenerative diseases (Arnold et al., 2018), and it has been shown that a diet consisting of high-calorie and processed foods increases the risk of depression and anxiety (Psaltopoulou et al., 2013; Lai et al., 2013). Although diet regimens were mostly associated with bad mood, irritability, anger resulting from food deprivation (Appleton & Baker, 2015), several studies found positive effects of fasting on mood and quality of life (Bowen et al., 2018; Hussin et al., 2013; Nugraha et al., 2020).

Many postprandial processes such as digestive absorption, glucose tolerance, and postprandial energy expenditure show diurnal oscillations, suggesting that the human metabolism is optimized for morning food intake (Ruddick-Collins et al., 2018). With the development of artificial lighting and the stressful Western lifestyle, people consume meals more often in the evening, which has been connected to desynchronization of internal clocks (Currenti et al., 2021). Research on night shift workers consistently shows that desynchronization has a number of negative outcomes, including reduced cognitive performance (Chellappa et al., 2018; Chellappa et al., 2019) and a negative impact on mood (Bedroisan & Nelson, 2017). Among other symptoms, impaired circadian rhythms result in daytime sleepiness and impaired sleep quality (Jafari Roodbandi et al., 2015), which has also been associated with poorer performance in cognitive tests (Lo et al., 2016; Nebes et al., 2009), fatigue, depression, anxiety, and confusion (Short & Louca, 2015). Time-restricted feeding (TRE), in which the eating cycle takes place over the course of the day, shows potential for rebalancing and aligning internal circadian clocks (Chaix et al., 2019).

Potential relations between gut-brain axis, intermittent fasting, and cognitive functions have also been discussed. The diversity of gut microbiome is essential for human health (Ceppa et al., 2018), and impaired gut microbiome was associated with a number of diseases such as

type 2 diabetes (Forslund et al., 2015), atopy in children (Fujimura et al., 2016), and autoimmune diseases (De Luca & Shoenfeld, 2019). Gut-brain axes microbiota is connected to the brain in a two-way communication that uses neural, endocrine, and immune signals. Previous research has found that mice growing in a sterile environment show reduced levels of the brain-derived neurotrophic factor in the hippocampal and cortical areas (Yamada et al., 2002). A recent study in rodents showed that intermittent fasting for 28 days increased the diversity of the microbiota and improved the performance of the Morris water maze task in diabetic mice. However, in the group that was previously treated with antibiotics (which disrupted the microbiota), such an improvement was absent (Liu et al., 2020).

The aforementioned mechanisms represent a theoretical background for postulating the effects of dietary regulation on cognitive functions and the overall psychological health. However, in most studies linking dietary regimens and cognitive functions, some form of caloric restriction was used as a dietary intervention. Research related to restricting the time period of food intake, rather than reducing calories, mostly refers to the alternate day fasting, or includes clinical populations or members of the Islamic religion during Ramadan fasting. Given the lack of research linking the TRE protocol to psychological functioning, the aim of this study was to investigate the potential benefits of TRE protocol with a more comprehensive approach that included a battery of cognitive tests (attention, memory, working memory, executive functions) as well as self-reported measures of mood and sleep quality, across three testing sessions (immediately prior to the start of fasting regimen, 1 month, and 2 months into the regimen) in a group of healthy volunteers adhering to TRE protocol, and a control group.

## Methods

### Study design

The present study was an open-label non-randomized study, with two groups of participants (experimental – adhering to the TRE regimen, and control – continuing with their regular diet) and a repeated measures design. While the lack of randomization can present a major methodological weakness, ensuring continued participants' motivation to both adhere to the TRE regimen and participate in continuing time points throughout the study, was central to the study goals. Therefore, participants were allowed to choose, based on their preferences, whether they wanted to adhere to the TRE regimen and participate in the testing sessions (experimental group) or just participate in the testing sessions without changing their dietary habits (control group) throughout the next 2 months. Participants were not given any financial or other incentives to participate in the study.

G*Power (Faul et al., 2007) analysis showed that in order to detect a group by time interaction for two groups being measured across three time points, an alpha of .05, power of .80, and small treatment effect of $f$ = .1, a total sample size of 98 participants is required. Due to the lack of any financial incentives, we expected a substantial dropout rate during the study, and opted to initially recruit at least twice as many participants.

### Participants

The study was conducted according to the Declaration of Helsinki and approved by the Institutional Review Board of Department of Psychology, Faculty of Humanities and Social Sciences in Zagreb. The participants were recruited via calls on social networks. The research was conducted in Croatia, with participants residing in different regions of the country. It was not mandatory for them to be in the capital city, as all the tests were completed online, and the invitations to participate in the research were also sent online. Out of the 411 volunteers who initially registered, 310 healthy adults were invited to participate. All selected participants were aged over 18, reported not being on any dietary regimen three months before the start of the study and did not meet any of the exclusion criteria. Participants' health status was determined in consultation with a general practitioner. The exclusion criteria were based on recommendations described by LeCheminant et al. (2013): Participants were not eligible to participate if they had cardiovascular diseases, diabetes, carcinoma, history of eating disorders, history of fainting caused by caloric

**Table 1** Physical characteristics of the participants

| | Experimental group (*n*=76) | | | | Control group (*n*=29) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *Min* | *Max* | *M* | *SD* | *Min* | *Max* | *t* | *p* | *d* |
| Age | 26.34 | 8.43 | 19 | 56 | 26.14 | 10.11 | 18 | 61 | 0.101 | .920 | 0.02 |
| Height (cm) | 172.22 | 8.93 | 150 | 196 | 170.50 | 8.38 | 155 | 186 | 0.884 | .379 | 0.19 |
| Body mass index ($kg/m^2$) | 24.06 | 2.97 | 19.27 | 33.79 | 22.60 | 2.68 | 18.78 | 30.39 | 0.022 | .982 | 0.50 |
| Lean body mass (The Boer Formula) | 51.20 | 6.67 | 36.51 | 73.39 | 48.93 | 5.88 | 39.38 | 61.10 | 1.585 | .116 | 0.35 |

restriction, anemia, kidney failure, used oral contraceptives, were pregnant at the time of the study, had a body-mass index (BMI) lower than 18.5, and/or had other acute and chronic conditions that deplete the patient's energy reserves. Following the informed consent, they received further instructions regarding participation in the study. A detailed participant flow diagram is shown in Figure 1.

The gender composition of the participants who remained until the final time point in the study did not differ between groups ($p < .01$, Fisher's exact test): There were 16 men and 60 women in the experimental group and seven men and 22 women in the control group. There were no significant differences between the control and experimental groups in age, height, and body mass index at the beginning of the study (see Table 1). Physical characteristics of the participants are shown in Table 1.

## Instruments

When applying for participation in the study via an online form, the participants reported their height and weight, from which their BMI was calculated. Four cognitive tests (objective measures of cognitive functioning) and two questionnaires (self-reported measures of sleeping quality and mood) were used at each testing session, as well as several follow-up questionnaires constructed ad hoc for the purposes of registering participants' adherence to the TRE regimen during the study. The questionnaires were administered via Google Forms, while the cognitive tasks were presented using E-Prime 3.0 software (Psychology Software Tools, Pittsburgh, PA).

*Attention*

The Attention Networking Test (Fan et al., 2002; McConnell & Shore, 2011) measures three aspects of attention: arousal (readiness to respond to a stimulus), directing attention (the selection of information from sensory input), and executive control (choosing between possible responses). After the initial instructions, the participants were shown arrows on the screen and their task was to decide the direction of the central arrow. If the arrows pointed to the left, they should have pressed the "1" button, if they pointed to the right, they should have pressed the "2" button. In the version used in this study arrows were presented in three uniform conditions that alternated by case: congruent (e.g. <<<<<), incongruent (e.g. >><>>), or neutral (e.g. 00<00). During the trials, participants were presented with a fixation dot (1000 ms) in the center of the screen (no cue condition). In addition to the above, during some trials an asterisk was briefly displayed on the screen. If an asterisk was presented above or below the fixation point (spatial cue), arrows also appeared above or below. When the asterisk appeared in the center of the screen (central cue), arrows were presented either above or below the fixation point. All conditions were counterbalanced for each participant. Participants first went through a cycle of 12 trials after which they received performance feedback. The three aspects of attention are calculated as follows: The arousal score is obtained by subtracting the average reaction time in the central cue condition from the average reac-

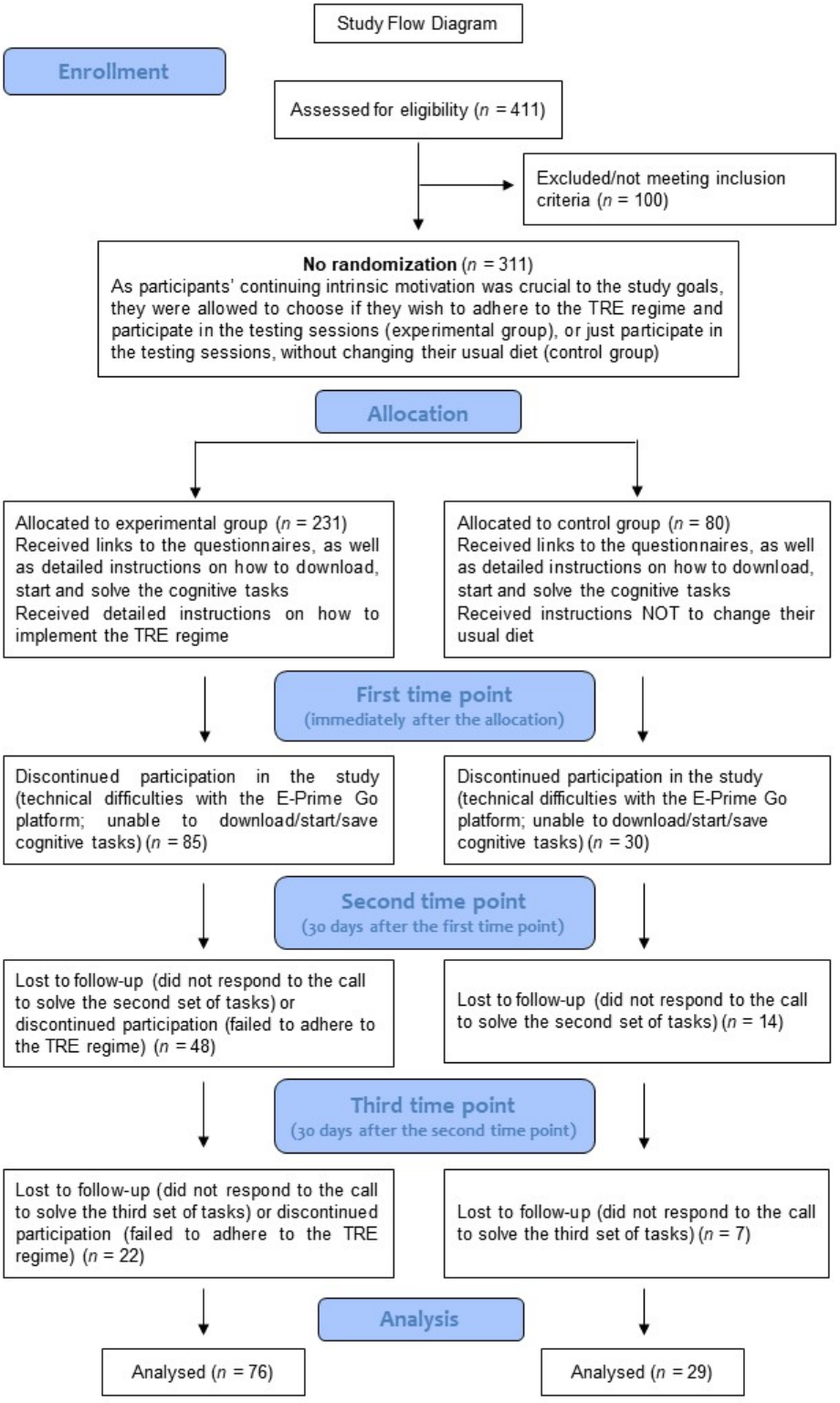



**Figure 1** Study Flow Diagram

tion time in the no-cue condition. The attentional focus score is obtained by subtracting the average reaction time in the spatial cue condition from the average reaction time in the central cue condition. Finally, the executive control score is obtained by subtracting the average reaction time in the congruent condition from the average reaction time in the incongruent condition (McConnell & Shore, 2011). In the present study, the test-retest correlations across three time points ranged between $r$ = .796 and $r$ = .878 for congruent conditions, between $r$ = .878 and $r$ = .906 for incongruent conditions, and between $r$ = .752 and $r$ = .859 for neutral conditions. Similarly, regarding conditions with different cues, the test-retest correlations ranged between $r$ = .801 and $r$ = .879 for conditions without cue, between $r$ = .824 and $r$ = .908 for spatial cue conditions, and between $r$ = .751 and $r$ = .879 for center cue conditions. As for scores extracted from the reaction times (RTs) across conditions, the test-retest correlations for the arousal and attention were low and ranged between $r$ = .247 and $r$ = .358 which is not unexpected, given the state and context dependent fluctuations in these functions. On the other hand, the executive control scores test-retest correlations ranged from $r$ =.691 to $r$ = 772.

*The memory span*

Two basic versions of memory tasks were used: Digit-Span Forward Task and Digit-Span Backward Task. In both versions, the participants are presented with a series of numbers, and their task is to repeat the presented numbers in the order in which they were presented (forwards) or in the reverse order of the presented ones (backwards; Bopp & Verhaeghen, 2005). These two versions of the task provide somewhat different measures: The forward memory version primarily measures attention efficiency and short-term memory capacity, while the backward memory version measures the working memory capacity, as it requires participants to recruit the central executor (Giofrè et al., 2016; Kasper et al., 2012). Such a view is confirmed by the fact that the two versions correlate differently with measures of intelligence (Cornoldi et al., 2013) and that different neural circuits are activated when solving different versions (Rossi et al., 2013). In the computer version of the task, participants were presented with a series of digits on the screen, starting from a minimum of three to a maximum of nine digits. After the presentation of each sequence, the participant's task was to use the keyboard to type the presented sequence on the screen in the same order or in the reverse order, according to the type of task. For each sequence of digits, the respondent had two attempts, whereby, if the respondent answered correctly both times, the number of digits increased by one (up to nine digits), and if they answered incorrectly both times, the test stopped. The longest sequence of digits that the examinee accurately reproduces in each of the tasks represents the scope of their short-term and working memory. The test-retest correlations across three time points for the digit span forward ranged between $r$ = .278 and $r$ = .371 – again, these low correlations are not surprising, since attention is highly variable as a function of function of context and testing conditions, which could not be controlled in this study. For the digit span backward, the test-retest correlations ranged between $r$ = .430 and $r$ = .451.

*Working memory*

The Fixed N-Back task, which was used in this study, involves a continuous sequence of stimuli (pictorial or graphic) presented gradually. The participant's task is to determine whether the current stimulus is the same as the previous one or the one before it. To successfully solve the task, it is necessary to activate a number of cognitive processes: coding and temporary storage of a series of stimuli and continuous updating of upcoming stimuli. At the same time, irrelevant stimuli should be inhibited and removed from working memory, while current stimuli should be timely compared with those currently held in memory (Rac-Lubashevsky & Kessler, 2016). The nature of the task requires the simultaneous recruitment of all the above-mentioned processes, which led to the classification of the N-back task among measures of working memory (Jaeggi et al., 2010). In our version of the task, the participants were shown letters on the screen, and they (considering the given rule) had to decide whether the presented letter was the target or not by clicking the "target" button. In the first condition, which served only to direct attention, only the letter Z was the target. In the 1-n back condition, a

letter was the target if it was equal to the last letter that appeared before it. In the 2-n back condition, a letter was the target if it was equal to the penultimate letter presented. Considering the simplicity of the 1-n back condition, we decided to use only the average reaction time in the 2-n back condition as a measure of working memory. The test-retest correlations across three time points for the N-back score ranged between $r$ = .767 and $r$ = .842.

*Executive functions*

The classic Stroop task examines the inhibitory control, which is the participant's ability to inhibit automatic responses and select relevant sensory information (Miller & Cohen, 2001). In the E-Prime 3.0. adaptation of the Stroop task, the participants were shown the names of colors (green, yellow, blue, red) on the screen, whereby the names were colored either in the same color as the written word (the word "yellow" colored in yellow) or in a different color (the word "yellow" colored blue), that is, the stimuli were congruent or incongruent. The participants' task was to determine in which color the word was written by pressing a predetermined key on the keyboard. Success in the task was determined as the average difference in reaction time to congruent and incongruent stimuli in milliseconds (ms), where a smaller difference indicates a greater inhibitory control. Participants first went through a cycle of eight trials, after which they received performance feedback. The test-retest correlations across three time points ranged between $r$ = .758 and $r$ = .838 for congruent conditions, and between $r$ = .815 and $r$ = .874 for incongruent conditions. The test-retest correlations for the Stroop score computed from these RTs ranged between $r$ = .589 and $r$ = .603.

*Sleep*

To measure sleep quality, the Pittsburgh Sleep Quality Index (PSQI; Buysse et al., 1989) was administered. This is a 19 items self-report questionnaire which aims to determine the participants' sleep quality on seven subscales: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleep medication, and daytime drowsiness. The scores range from 0 to 21 and the authors suggest that a score > 5 is to be considered a significant sleep disturbance. The reliability reported in earlier studies was between α =.70 and α =.83 (Mollayeva et al., 2016). In this study, McDonald's ω varied across time points between ω = .63 and ω = .68.

*Mood*

The Adjective Check List (ACL; Taub & Berger 1974; Croatian translation and validation: Radošević-Vidaček et al., 1990), which consists of 57 adjectives that denote different emotional states, was used in this study. Participants had to respond to what extent they experienced certain emotions that day, on a scale ranging from 0 ("Not at all") to 4 ("Extremely"). Average scores were computed for eight dimensions: anxiety (six items), depression (six items), fatigue (eight items), hostility (eight items), friendliness (five items), cheerfulness (five items), concentration (eight items), and energy (six items). The internal reliability coefficients of the subscales were high, ranging from ω = .78 for concentration to ω = .92 for fatigue.

## Procedure

The study was conducted over a period of 2 months (May-July 2021) online, via the Google Forms for the questionnaires and E-PrimeGo (version 3.0) platform for the cognitive tests. After filling in the initial questionnaire, the participants received an e-mail confirming or rejecting their participation in the study with a detailed explanation. All participants received instructions for solving the tasks and filling in the questionnaires prior to the first testing session. Participants could access the E-PrimeGo test base using a link provided by the researcher. Their answers were automatically saved in an online database accessible only to the researchers. Video instructions on the installation of the tasks were sent to the participants, as well as instructions for solving potential technical difficulties. The participants were given 3 days to solve all the tests of the first session. They were instructed to start the session only if they were able to do so in a quiet environment without any distractions.

After the completion of the first testing session, participants in the experimental group received detailed instructions for starting TRE. They were introduced to the concept of TRE,

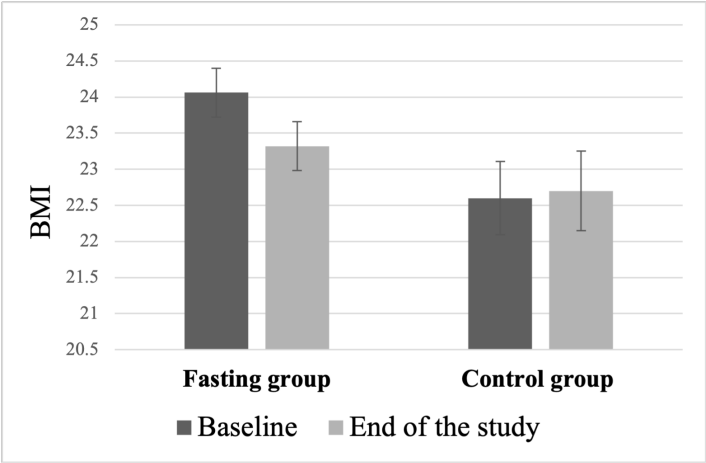



**Figure 2** BMI at the baseline and at the end of the study (following two months of intermittent fasting regime for the experimental group and no dietary changes in the control group). Error bars represent standard errors.

and its benefits and risks. Along came the instruction that in the next two months they should adhere to the TRE, for which they alternate between a 16-hour fasting window and an 8-hour eating window. Participants were given the choice to decide when to start their fasting/eating window, with the instruction that they maintain a routine (e.g. if they decide to start their fasting window at 5pm, it is important that they maintain that routine every day). The participants were encouraged to not drastically change the type of food they consume and to pay attention to the intake of enough nutrients in the eating window. During the fasting window, participants were allowed to consume water, unsweetened tea and coffee. In the first week, participants were allowed to maintain a fasting window of 14 hours and an eating window of 10 hours to facilitate the participants' transition to a full fast of 16 hours. Also, we encouraged the subjects to take one day a week "off" from fasting, during which they would not pay attention to the time intervals of eating and fasting. Participants from the control group were instructed not to adhere to any new diet during the participation in the study, but they were included in all the same consecutive testing sessions as the participants from the experimental group.

The second testing session took place 1 month after the first session, and the third session took place 2 months after the first one. Due to an online participation and different work/life schedules of participants, there was no fixed time of day when participants were instructed to do the testing. The testing announcements were sent out in the morning exactly a month (and two months) after the first testing took part, and there was an instruction to finish all the tests and fill out the questionnaires in 3 days. Participants completed all cognitive tests and filled the mood and sleep questionnaires, as well as control questionnaires about weight loss and the adherence to the TRE protocol.

Throughout the study, we maintained regular contact with the participants in order to motivate them to stay in the dietary regimen and keep completing the cognitive tests, and to address any of their doubts and questions. Participants from the experimental group could join a private group on Facebook, in which tips for more efficient adherence to fasting and answers to the most common questions were published, and the participants themselves could comment and start discussions. All contents published in the group have been reviewed and approved by a MSc nutritionist, who also held an online lecture for the participants in order to familiarize them with intermittent fasting in more detail. A summary and the most important tips were sent to the e-mail addresses of all participants who could not participate in the online lecture.

### Statistical analyses

For each dependent variable, we conducted an analysis of variance for repeated measures with the group (experimental vs. control) as a source of variance between participants and time-point ($1^{st}/2^{nd}/3^{rd}$) as a within-group source of variance. In instances where Mauchly's test of sphericity showed that variances between groups differed, Greenhouse-Giesser correction was applied.

## Results

### Weight loss

Although weight loss was not a primary focus of this study, we considered the changes in

BMI as objective indicators of successful adherence to the intermittent fasting regimen. The repeated measures ANOVA showed a significant time x group interaction, $F(1, 102) = 31.32$, $p < .001$, $\eta_p^2 = .235$. As can be seen in Figure 2, the interaction stemmed from the fact that the average BMI in the experimental group was lower at the end of the study compared to the baseline, $t(75) = 8.86$; $p < .001$; $d = 0.25$, while the average BMI in the control group did not change significantly, $t(27) = -1.11$; $p = .27$; $d = 0.03$.

### Performance on cognitive tests

As can be seen in Table 2, there was a significant increase in scores on the executive control segment of the Attention Networking Test, digit span forward, N-back, and Stroop as a function of time point. However, there were no significant differences between the experimental and the control group, and no significant group x time-point interactions, which implies that these improvements were a result of practice and familiarity with the task, and not the TRE regimen per se.

### Self-report measures

As can be deduced from Table 3, there were several significant differences in mood dimensions as a function of the time point: Participants scored lower on anxiety, depression, and fatigue, and higher on cheerfulness and energy at the end of the study, compared to the pre-intervention testing. However, as in the case of cognitive tests, there were no differences between groups, and no group x time point interactions, implying that these shifts in mood cannot be ascribed to the TRE regimen either.

## Discussion

The aim of this study was to determine how TRE affects cognitive functions and mental health, that is, whether fasting for 16 hours a day throughout two months would lead to significant improvements in attention, memory, executive skills, mood, and sleep quality.

### The impact of TRE on cognitive functions.

We found no effect of TRE on cognitive performance. Scores on several cognitive tasks improved significantly across time points, but for both groups. Given that there were no significant group x time point interactions, it is most likely that the observed trend was caused by familiarity with the tasks and solving strategies (Wesnes & Pincock, 2002). Previous studies reported mixed findings, with some reporting positive effects of intermittent fasting on cognitive measures (Giles et al., 2012; Colzato et al., 2013; Farooq et al., 2010; 2015; Teong et al., 2021), while others found no effect (Ghayour Najafabadi et al., 2015; Harder et al., 2017; Rachid et al., 2021), or even negative effects, at least in the case of short-term fasting (Benau et al., 2014; 2021). There might be numerous methodological reasons we failed to observe significant change in cognitive functioning of the fasting group (see limitations section for more details), and the lack of evidence does not prove that an effect does not exist. Nonetheless, it is still worth noting that across seven measures in three domains of cognitive functioning, none showed a significant group x time interaction. Thus, there is a possibility that the positive effects of intermittent fasting on domains beyond weight loss are overemphasized in the media and popularized without adequate evidence (Johnstone, 2014). Caution might prove especially important in cases where there is a risk that fasting could be a sort of gateway to extreme food restrictions and other harmful behaviors. It has already been shown that intermittent fasting is related to eating disorder behaviors and psychopathology (Ganson et al., 2022).

The most researched dietary intervention in humans is still caloric restriction, and conclusions are then generalized to similar dietary regimens such as intermittent fasting. Although caloric restriction and intermittent fasting have similar effects (Teong et al., 2021), it should be emphasized that intermittent fasting does not necessarily limit the amount of food consumed. Some studies showed that fat tissue loss can occur even without caloric restriction (Moro et al., 2016). Furthermore, it seems that the type of consumed food also has an impact on cognitive functions. Thus, for example, a diet with a high fat content is associated with worse performance in cognitive tasks (Edwards et al., 2011). In a recent systematic review, it has been concluded that the observed

**Table 2** Scores on cognitive tasks (attention, memory, executive functions) across time points for experimental and control group, and the results of repeated measures ANOVAs, with group (experimental vs. control) as a source of variance between participants and time point (1st/2nd/3rd) as a within-group source of variance.

| | First time point | | Second time point | | Third time point | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Expmtl.** | **Control** | **Expmtl.** | **Control** | **Expmtl.** | **Control** | | | | |
| | *M (SD)* | *M (SD)* | *M (SD)* | *M (SD)* | *M (SD)* | *M (SD)* | | *F* | *p* | $\eta_p^2$ |
| **Attention** | | | | | | | | | | |
| Arousal | 14.70 (32.05) | 14.61 (20.65) | 14.84 (21.29) | 15.11 (23.69) | 9.62 (26.25) | 13.48 (21.26) | Time point | 0.39 | .67 | .004 |
| | | | | | | | Group | 0.18 | .65 | .002 |
| | | | | | | | Time point x group | 0.13 | .88 | .001 |
| Directing attention | 23.05 (31.98) | 12.04 (26.64) | 23.39 (32.02) | 14.70 (25.75) | 22.67 (28.12) | 14.95 (24.13) | Time point | 0.07 | .93 | .001 |
| | | | | | | | Group | 4.08 | .05 | .041 |
| | | | | | | | Time point x group | 0.08 | .93 | .001 |
| Executive control | 129.16 (87.35) | 128.38 (60.92) | 106.08 (76.82) | 99.10 (37.33) | 94.36 (48.42) | 83.55 (27.25) | Time point | 21.92 | .001 | .187 |
| | | | | | | | Group | 0.21 | .65 | .002 |
| | | | | | | | Time point x group | 0.34 | .71 | .004 |
| **Memory** | | | | | | | | | | |
| Digit span forward | 7.24 (0.96) | 6.81 (1.08) | 7.45 (1.03) | 7.30 (1.07) | 7.48 (1.04) | 7.41 (1.05) | Time point | 5.15 | .007 | .051 |
| | | | | | | | Group | 1.68 | .42 | .017 |
| | | | | | | | Time point x group | 0.88 | .20 | .009 |
| Digit span backward | 6.13 (1.40) | 6.08 (1.04) | 6.37 (1.38) | 6.32 (1.28) | 6.56 (1.23) | 6.40 (1.04) | Time point | 2.24 | .11 | .024 |
| | | | | | | | Group | 0.16 | .69 | .002 |
| | | | | | | | Time point x group | 0.06 | .94 | .001 |
| N-back | 1.75 (0.93) | 1.81 (0.83) | 1.95 (0.82) | 1.64 (0.89) | 2.04 (0.96) | 2.20 (0.97) | Time point | 9.07 | .001 | .092 |
| | | | | | | | Group | 0.33 | .57 | .004 |
| | | | | | | | Time point x group | 0.70 | .49 | .008 |
| **Executive functions** | | | | | | | | | | |
| Stroop test | 168.78 (151.71) | 192.05 (124.63) | 124.28 (133.13) | 114.17 (78.73) | 80.12 (82.12) | 68.89 (68.34) | Time point | 32.11 | .001 | .251 |
| | | | | | | | Group | 0.01 | .98 | .000 |
| | | | | | | | Time point x group | 1.09 | .34 | .011 |

cognitive benefits were associated more with other interventions, such as decreasing the caloric intake without a complete fasting period, weight loss, dietary approaches to stop hypertension, blood pressure reduction, and exercise, rather than with intermittent fasting per se (Senderovich et al., 2023).

One possible explanation for the lack of effect found in this study, is the low level of control the researchers had over the type and amount of food consumed, and the partcipants' weight measurement. Since in our study we did not control the type and amount of food consumed, and the participants measured their own weight, greater control over the aforementioned factors is recommended in future studies. Another possible explanation for the lack of effect is that the adherence to the diet in our sample was not strict enough. Participants who did not fast for at least 14 hours a day throughout the duration of the study (based on their self-reported adherence) were excluded from the final analyses, but a greater control of time periods and the number of days spent fasting is needed. That being said, we have reason to believe that our participants were highly intrinsically motivated and that they did adhere to the time restricted eating regime to the best of their possibilities: 71% of participants from the experimental group reported that they planned to continue with the intermittent fasting even after the completion

**Table 3** Scores on mood and sleep quality questionnaires across time points for experimental and control group, and the results of mixed model repeated measures ANOVAs, with group (experimental vs. control) as a source of variance between participants and time point (1st/2nd/3rd) as a within-group source of variance.

| | First time point | | Second time point | | Third time point | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Expmtl. | Control | Expmtl. | Control | Expmtl. | Control | | | | |
| | *M (SD)* | *M (SD)* | *M (SD)* | *M (SD)* | *M (SD)* | *M (SD)* | | *F* | *p* | $\eta_p^2$ |
| Anxiety | 0.84 (0.80) | 0.87 (0.77) | 0.95 (0.96) | 0.99 (0.85) | 0.63 (0.78) | 0.71 (0.62) | Time point | 4.66 | .01 | .045 |
| | | | | | | | Group | 0.17 | .68 | .002 |
| | | | | | | | Time point x group | 0.01 | .99 | .000 |
| Depression | 0.86 (0.81) | 0.95 (0.75) | 0.84 (0.86) | 0.90 (0.67) | 0.62 (0.76) | 0.65 (0.63) | Time point | 4.72 | .01 | .045 |
| | | | | | | | Group | 0.37 | .54 | .004 |
| | | | | | | | Time point x group | 0.07 | .94 | .001 |
| Fatigue | 1.42 (0.98) | 1.64 (0.97) | 1.30 (0.98) | 1.87 (1.03) | 1.24 (0.96) | 1.35 (1.09) | Time point | 3.21 | .04 | .031 |
| | | | | | | | Group | 2.74 | .10 | .027 |
| | | | | | | | Time point x group | 1.86 | .16 | .018 |
| Hostility | 0.47 (0.54) | 0.50 (0.67) | 0.45 (0.68) | 0.52 (0.69) | 0.38 (0.57) | 0.37 (0.49) | Time point | 2.03 | .13 | .020 |
| | | | | | | | Group | 0.09 | .76 | .001 |
| | | | | | | | Time point x group | 0.42 | .65 | .004 |
| Friendliness | 2.44 (0.97) | 2.53 (0.91) | 2.55 (0.91) | 2.45 (0.98) | 2.64 (0.96) | 2.73 (0.98) | Time point | 2.34 | .09 | .023 |
| | | | | | | | Group | 0.01 | .95 | .000 |
| | | | | | | | Time point x group | 0.46 | .64 | .005 |
| Cheerfulness | 2.11 (1.00) | 2.24 (0.66) | 2.26 (0.95) | 2.14 (0.82) | 2.43 (0.92) | 2.54 (0.96) | Time point | 5.03 | .007 | .048 |
| | | | | | | | Group | 0.01 | .98 | .000 |
| | | | | | | | Time point x group | 0.69 | .50 | .007 |
| Concentration | 2.40 (0.76) | 2.48 (0.70) | 2.58 (0.75) | 2.35 (0.62) | 2.63 (0.78) | 2.56 (0.84) | Time point | 1.36 | .26 | .014 |
| | | | | | | | Group | 0.38 | .54 | .004 |
| | | | | | | | Time point x group | 1.12 | .33 | .011 |
| Energy | 1.75 (0.93) | 1.81 (0.83) | 1.95 (0.82) | 1.64 (0.89) | 2.04 (0.96) | 2.20 (0.97) | Time point | 6.04 | .003 | .058 |
| | | | | | | | Group | 0.05 | .82 | .001 |
| | | | | | | | Time point x group | 2.16 | .14 | .021 |
| Sleep quality (PSQI) | 5.21 (3.23) | 4.44 (1.95) | 4.00 (2.61) | 4.67 (2.76) | 4.72 (2.76) | 4.72 (2.76) | Time point | 0.76 | .47 | .015 |
| | | | | | | | Group | 0.02 | .89 | .000 |
| | | | | | | | Time point x group | 0.89 | .14 | .040 |

of the study. This is in line with the claims that this particular dietary regime is easy to implement and adhere to, even in the long run. In fact, it might be argued that its implementation becomes even easier as a function of time, as it has previously been shown that contrary to the effects of short-term, selective food deprivation, long-term energy restriction decreases food cravings (Meule, 2020).

### Impact of intermittent fasting on mood and sleep quality.

We also examined whether the participants' mood and sleep quality changed in the context of intermittent fasting. Again, the only significant effects were related to the time point: Participants scored lower on anxiety, depression, and fatigue, and higher on cheerfulness and energy at the end of the study, compared to the pre-intervention testing. However, there were no differences between groups, or group

x time point interactions in either of the mood dimensions, or in sleep quality. Given that the research was conducted in the period from May to July, the positive changes in the mood of both groups could be linked to the arrival of Summer months and the positive effects of sunny weather on mood (Keller et al., 2005). Summer vacations might have also influenced this increase in positive and decrease in negative moods.

Our findings are in line with some previous studies, showing that that neither a short-term two-day fast nor an intermittent fast lasting 8 weeks has any effect on the mood of the participants or on the quality of sleep (Solianik & Sujeta, 2018; Teong et al., 2021). The assumptions about the positive influence of fasting on mood are based primarily on studies of the clinical population, where dietary interventions have shown significant potential in improving the symptoms of mood disorders in obese participants with frequent comorbidities (Patsalos et al., 2021). Although some studies showed that intermittent fasting might have potential for improving mood (Bowen et al., 2018) and reducing depressive symptoms (Hussin et al., 2013) in non-clinical populations, it is possible that the same dietary interventions have greater implications for clinical populations. Additionally, gender differences were found in the impact of fasting on the reduction of anxiety (with a significant reduction [illegible] observed in men only; Nu
and due to the small numb
pants, we could not addre
this study. Finally, different
and different fasting regim
various studies, so due to t
designs using TRE, only limit
be drawn.

Recently, it has been su
mittent fasting has a posit
through the regulation of
(Chaix et al., 2019). We wer
any significant effect of fasti
although the results did sho
of improvement in PSQI s
fasting group had the ave
above 5, which is clinically
paired sleep quality, dropp
second time point). In futu
it would be interesting to cl
ity of sleep has a mediating influence on the relationship between intermittent fasting and performance on cognitive tasks, considering that there is no research that connects the mentioned variables into one complete model.

### Limitations of the current study and recommendations for future research

The major limitation of this study was the lack of randomization. As previously described, due to the nature of the research topic, and the fact that the participants did not receive any financial incentives, their intrinsic motivation was the only factor ensuring their continuing cooperation (both in adhering to the TRE, and in participating in testing sessions). Thus, their preferences were accommodated, at the cost of internal validity of the study. Similarly, there was no control over the exact time of day when the participants started fasting, which might be crucial, as it has been speculated that intermittent fasting has an impact on circadian rhythms. The participants in this study were instructed to start the fasting period in the evening, because interventions with reduced food intake in the evening have been shown to be more effective (Jakubowicz et al., 2013), but due to differences in the schedule and working hours of the participants, they were allowed to decide for themselves when to start with the fasting

## Original purpose

Intermittent fasting drew attention among scientists in fields outside of nutrition, due to reports of its positive effects on psychological well-being. However, in most studies that link dietary regimens to cognitive functions, some form of caloric restriction was used as a dietary intervention. Given the lack of research linking intermittent fasting and a person's psychological health and cognitive functioning, the goal of this study was to contribute to the literature investigating the potential benefits of intermittent fasting, with a more comprehensive approach. This included a battery of cognitive tests (objective measures of cognitive functioning) as well as reports of mood (subjective measures of psychological functioning) and sleep quality, as it has been suggested that intermittent fasting regulates circadian rhythms and thus affects psychological outcomes.

period. Eating food late in the evening can disrupt circadian rhythms and consequently performance in cognitive tests (Currenti et al., 2021). Future studies should equalize fasting conditions for all subjects, and/or consider the interindividual differences in morningness - eveningness in order to determine whether there are differences in time-restricted day versus night interventions. Related to this, another potentially important element we could not control for was the time of day when participants took the tests and filled the questionnaires – both circadian rhythms and timing in relation to fasting window could have influenced the performance on tests. As we argued above, we opted to achieve higher ecological validity, aiming to assess the outcomes of the type of time restricted eating regime people actually implement in real life, not laboratory conditions, though this comes at a cost of internal validity.

Additionally, online testing - as opposed to testing in a lab setting - could have impaired the validity of the results. Solving the tests without the presence of the researcher could have led to uneven conditions among the participants, starting from physical surroundings, the presence of other people in the room, to interrupting the testing and solving the tests at different intervals.

Last, but not least, our control group was disproportionately smaller than the fasting group, and by the end of the study, fewer than 30 participants remained in it. This was an oversight on our part: We had expected a large dropout in the experimental group (because adherence to any fasting regime is challenging) and minimal dropout in the control group (because they only had to solve the tasks and fill the questionnaires on three occasions – all online, no dietary or any other lifestyle changes were required). Ultimately, however, the dropout in both groups was large. We are very grateful to one of the reviewers of this paper (Dr. Sean Devine) who pointed out that imbalanced sample sizes between groups decrease power in repeated-measures designs. He also took the time to conduct a simulation-based power analysis using multilevel modeling to account for the uneven sample sizes. Depending on the expected effect size he used, he found that the current dataset would only have power to detect a small interaction effect size ($\beta = .1$ or $\beta = .2$) 17% to 50% of the time, respectively. This implies that the study was very likely underpowered to detect the effects of these magnitudes. While we agree whole-heartedly with the fact that this sample size would not enable us to detect such small effects, we also have to draw the readers' attention to the main point of this manuscript, which is calling for the caution in advising dietary restrictions and taking into account costs and benefits of such restrictions. We used the smallest possible effects size in our calculations of the statistical power: Even if the effect was detected, future studies might try to assess the applicability of such small effects in real-life, especially when therapeutic benefits are expected, and even more so when the interventions are not without potential adverse effects.

## Conclusion

Questions ranging from seemingly simple ones, such as what best defines dieting or restrained eating to the methodologically complex ones, such as which approaches to the assessment of eating behaviors are optimal, are often a matter of debate even among the experts (for a detailed review see Meule, 2023). To laymen, especially those in search of solutions for their physical or mental health related problems, setting realistic expectations and even more so assessing the effectiveness of a dietary intervention on themselves beyond the objective measure of weight loss is necessarily distorted by the various processes of motivated reasoning. Using a set of objective cognitive tasks, we found no substantial effect of the intermittent fasting on either the tested cognitive functions (attention, memory, working memory, executive functions), throughout the time period of two months. Similarly, we found no substantial effect of fasting on any of the dimensions of mood (anxiety, depression, fatigue, hostility, friendliness, cheerfulness, concentration, energy) or on sleep quality. Although the mechanisms underlying the impact of fasting on cognitive functions have been described in detail and show potential in animal models (Dias et al., 2021), research on humans is very limited.

Variational research designs represent an additional problem in drawing conclusions. Researchers have used different fasting regimens, where very often all regimens are classified un-

der a common denominator, although the fasting period takes place at different times and are of different durations. Ramadan fasting, of which effects on cognition are most often observed, takes place during the day, so it is not directly comparable to other fasting regimens in which the fasting period takes place at night. With the given body of evidence, it seems that when it comes to recommending fasting dietary regimens, greater care should be put into avoiding exaggerated claims about the benefits which transcend those related to physical health, primarily regulation of weight.

## Funding

The license for the E-Prime Go v3 was funded by the internal grant no. 11-929-1027 (Faculty of Humanities and Social Sciences). No other segments of the study were funded.

## Acknowledgments

The authors would like to thank Anja Bašnec, a M.Sc. nutritionist, affiliated with Food technology faculty in Osijek, J.J.S. University, Croatia, for her generous help with preparing the educational materials for the participants in the experimental group.

## References

Appleton, K. M., & Baker, S. (2015). Distraction, not hunger, is associated with lower mood and lower perceived work performance on fast compared to non-fast days during intermittent fasting. *Journal of Health Psychology, 20*(6), 702–711. https://doi.org/10.1177/1359105315573430

Arnold, S. E., Arvanitakis, Z., Macauley-Rambach, S. L., Koenig, A. M., Wang, H.-Y., Ahima, R. S., Craft, S., Gandy, S., Buettner, C., Stoeckel, L. E., Holtzman, D. M., & Nathan, D. M. (2018). Brain insulin resistance in type 2 diabetes and Alzheimer disease: Concepts and conundrums. *Nature Reviews Neurology, 14*(3), 168–181. https://doi.org/10.1038/nrneurol.2017.185

Bedrosian, T. A., & Nelson, R. J. (2017). Timing of light exposure affects mood and brain circuits. *Translational Psychiatry*, *7*(1), e1017-e1017. https://doi.org/10.1038/tp.2016.262

Benau, E. M., Makara, A., Orloff, N. C., Benner, E., Serpell, L., & Timko, C. A. (2021). How does fasting affect cognition? An updated systematic review (2013-2020). *Current Nutrition Reports*, *10*(4), 376–390. https://doi.org/10.1007/s13668-021-00370-4

Benau, E. M., Orloff, N. C., Janke, E. A., Serpell, L., & Timko, C. A. (2014). A systematic review of the effects of experimental fasting on cognition. *Appetite*, *77*, 52–61. https://doi.org/10.1016/j.appet.2014.02.014

Bopp, K. L., & Verhaeghen, P. (2005). Aging and verbal memory span: A meta-analysis. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *60*(5), P223-P233. https://doi.org/10.1093/geronb/60.5.p223

Bowen, J., Brindal, E., James-Martin, G., & Noakes, M. (2018). Randomized trial of a high protein, partial meal replacement program with or without alternate day fasting: Similar effects on weight loss, retention status, nutritional, metabolic, and behavioral outcomes. *Nutrients*, *10*(9), Article 1145. https://doi.org/10.3390/nu10091145

Buysse, D. J., Reynolds III, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Research*, *28*(2), 193-213. https://doi.org/10.1016/0165-1781(89)90047-4

Ceppa, F., Mancini, A., & Tuohy, K. (2018). Current evidence linking diet to gut microbiota and brain development and function. *International Journal of Food Sciences and Nutrition, 70*(1), 1–19. https://doi.org/10.1080/09637486.2018.1462309

Chaix, A., Lin, T., Le, H. D., Chang, M. W., & Panda, S. (2019). Time-restricted feeding prevents obesity and metabolic syndrome in mice lacking a circadian clock. *Cell Metabolism*, *29*(2), 303-319. https://doi.org/10.1016/j.cmet.2018.08.004

Chellappa, S. L., Morris, C. J., & Scheer, F. A. J. L. (2018). Daily circadian misalignment impairs human cognitive performance task-dependently. *Scientific Reports*, *8*, Article 3041. https://doi.org/10.1038/s41598-018-20707-4

Chellappa, S. L., Morris, C. J., & Scheer, F. A. J. L. (2019). Effects of circadian misalignment on cognition in chronic shift workers. *Scientific Reports, 9*, Article 699. https://doi.org/10.1038/s41598-018-36762-w

Colzato, L., Jongkees, B., Sellaro, R., & Hommel, B. (2013). Working memory reloaded:

Tyrosine repletes updating in the N-back task. *Frontiers in Behavioral Neuroscience*, *7*, Article 200. https://doi.org/10.3389/fnbeh.2013.00200

Cornoldi, C., Orsini, A., Cianci, L., Giofrè, D., & Pezzuti, L. (2013). Intelligence and working memory control: Evidence from the WISC-IV administration to Italian children. *Learning and Individual Differences*, *26*, 9-14. https://doi.org/10.1016/j.lindif.2013.04.005

Currenti, W., Godos, J., Castellano, S., Mogavero, M. P., Ferri, R., Caraci, F., Grosso, G., & Galvano, F. (2021). Time restricted feeding and mental health: A review of possible mechanisms on affective and cognitive disorders. *International Journal of Food Sciences and Nutrition*, *72*(6), 723-733. https://doi.org/10.1080/09637486.2020.1866504

De Cabo, R., & Mattson, M. P. (2019). Effects of intermittent fasting on health, aging, and disease. *New England Journal of Medicine, 381*(26), 2541–2551. https://doi.org/10.1056/nejmra1905136

De Luca, F., & Shoenfeld, Y. (2019). The microbiome in autoimmune diseases. *Clinical & Experimental Immunology*, *195*(1)*,* 74-85. https://doi.org/10.1111/cei.13158

Dias, G. P., Murphy, T., Stangl, D., Ahmet, S., Morisse, B., Nix, A., Aimone, L. J., Aimone, J. B., Kuro-O, M., Gage, F. H., & Thuret, S. (2021). Intermittent fasting enhances long-term memory consolidation, adult hippocampal neurogenesis, and expression of longevity gene Klotho. *Molecular Psychiatry*, *26,* 6365–6379. https://doi.org/10.1038/s41380-021-01102-4

Donnadieu-Rigole, H., Olive, L., Nalpas, B., Duny, Y., Nocca, D., & Perney, P. (2016). Prevalence of psychoactive substance consumption in people with obesity. *Substance Use & Misuse*, *51*(12), 1649-1654. https://doi.org/10.1080/10826084.2016.1191514

Edwards, L. M., Murray, A. J., Holloway, C. J., Carter, E. E., Kemp, G. J., Codreanu, I., Brooker, H., Tyler, D. J., Robbins, P. A., & Clarke, K. (2011). Short-term consumption of a high-fat diet impairs whole-body efficiency and cognitive function in sedentary men. *The FASEB Journal*, *25*(3), 1088-1096. https:/ / doi.org/ 10.1096/ fj.10-171983

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience, 14*(3), 340-347. https://doi.org/10.1162/089892902317361886

Farooq, A., Herrera, C. P., Almudahka, F., & Mansour, R. (2015). A prospective study of the physiological and neurobehavioral effects of Ramadan fasting in preteen and teenage boys. *Journal of the Academy of Nutrition and Dietetics*, *115*(6), 889-897. https://doi.org/10.1016/j.jand.2015.02.012

Farooq, S., Nazar, Z., Akhtar, J., Irfan, M., Subhan, F., Ahmed, Z., Khan, I. H., & Naeem, F. (2010). Effect of fasting during Ramadan on serum lithium level and mental state in bipolar affective disorder. *International Clinical Psychopharmacology*, *25*(6), 323-327. https://doi.org/10.1097/yic.0b013e3283466ed3

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175-191. https://doi.org/10.3758/BF03193146

Fernando, H. A., Zibellini, J., Harris, R. A., Seimon, R. V., & Sainsbury, A. (2019). Effect of Ramadan fasting on weight and body composition in healthy non-athlete adults: A systematic review and meta-analysis. *Nutrients*, *11*(2), Article 478. https://doi.org/10.3390/nu11020478

Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., Prifti, E., Vieira-Silva, V., Gudmundsdottir, V., Pedersen, H. K., Arumugam, M., Kristiansen, K., Voigt, A. Y., Vestergaard, H., Hercog, R., Costea, P. I., Kultima, J. R., Li, J., Jørgensen, T., … & Pedersen, O (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature, 528*, 262–266. https://doi.org/10.1038/nature15766

Fujimura, K. E., Sitarik, A. R., Havstad, S., Lin, D. L., Levan, S., Fadrosh, D., Panzer, A. R., LaMere, B., Rackaityte, E., Lukacs, N. V., Wegienka, G., Boushey, H. A., Ownby, D. R., Zoratti, E. M., Levin, A. M., Johnson, C. C., & Lynch, S. V. (2016). Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nature Medicine*, *22*(10), 1187-1191. https://doi.org/10.1038/nm.4176

Ganson, K. T., Cuccolo, K., Hallward, L., & Nagata, J. M. (2022). Intermittent fasting: Describing engagement and associations with eating disorder behaviors and psychopathology among Canadian adolescents and young

adults. *Eating Behaviors*, *47*, Article 101681. https://doi.org/10.1016/j.eatbeh.2022.101681

Ghayour Najafabadi, M., Rahbar Nikoukar, L., Memari, A., Ekhtiari, H., & Beygi, S. (2015). Does Ramadan fasting adversely affect cognitive function in young females? *Scientifica*, *2015*, Article 432428. https://doi.org/10.1155/2015/432428

Giles, G. E., Mahoney, C. R., Brunyé, T. T., Gardony, A. L., Taylor, H. A., & Kanarek, R. B. (2012). Differential cognitive effects of energy drink ingredients: Caffeine, taurine, and glucose. *Pharmacology Biochemistry and Behavior*, *102*(4), 569-577. https://doi.org/10.1016/j.pbb.2012.07.004

Giofrè, D., Stoppa, E., Ferioli, P., Pezzuti, L., & Cornoldi, C. (2016). Forward and backward digit span difficulties in children with specific learning disorder. *Journal of Clinical and Experimental Neuropsychology, 38*(4), 478–486. https://doi.org/10.1080/13803395.2015.1125454

Harder-Lauridsen, N. M., Rosenberg, A., Benatti, F. B., Damm, J. A., Thomsen, C., Mortensen, E. L., Pedersen, B. K., & Krogh-Madsen, R. (2017). Ramadan model of intermittent fasting for 28 d had no major effect on body composition, glucose metabolism, or cognitive functions in healthy lean men. *Nutrition, 37*, 92–103. https://doi.org/10.1016/j.nut.2016.12.015

Harvie, M. N., Pegington, M., Mattson, M. P., Frystyk, J., Dillon, B., Evans, G., Cuzick, J., Jebb, S. A., Martin, B., Cutler, R. G., Son, T. G., Maudsley, S., Carlson, O. D., Egan, J. M., Flyvbjerg, A., & Howell, A. (2010). The effects of intermittent or continuous energy restriction on weight loss and metabolic disease risk markers: A randomized trial in young overweight women. *International Journal of Obesity, 35*(5), 714–727. https://doi.org/10.1038/ijo.2010.171

Heilbronn, L. K., de Jonge, L., Frisard, M.I., DeLany, J.P., Larson-Meyer, D.E., Rood, J., Nguyen, T., Martin, C. K., Volaufova, J., Most, M. M., Greenway, F. L., Smith, S. R., Deutsch, W. A., Williamson, D. A., & Ravussin, E. (2006). Effect of 6-month calorie restriction on biomarkers of longevity, metabolic adaptation, and oxidative stress in overweight individuals: A randomized controlled trial. *JAMA*, *295*(13), 1539-1548. https://doi.org/10.1001/jama.295.13.1539

Hussin, N. M., Shahar, S., Teng, N. I. M. F., Ngah, W. Z. W., & Das, S. K. (2013). Efficacy of fasting and calorie restriction (FCR) on mood and depression among ageing men. *The Journal of Nutrition, Health & Aging, 17*(8), 674–680. https://doi.org/10.1007/s12603-013-0344-9

Idrobo, F., Nandy, K., Mostofsky, D. I., Blatt, L., & Nandy, L. (1987). Dietary restriction: Effects on radial maze learning and lipofuscin pigment deposition in the hippocampus and frontal cortex. *Archives of Gerontology and Geriatrics*, *6*(4), 355-362. https:/ / doi.org/ 10.1016/ 0167-4943(87)90014-8

Ingram, D. K., Weindruch, R., Spangler, E. L., Freeman, J. R., & Walford, R. L. (1987). Dietary restriction benefits learning and motor performance of aged mice. *Journal of Gerontology*, *42*(1), 78-81. https://doi.org/10.1093/geronj/42.1.78

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory, 18*(4), 394–412. https://doi.org/10.1080/09658211003702171

Jafari Roodbandi, A., Choobineh, A., & Daneshvar, S. (2015). Relationship between circadian rhythm amplitude and stability with sleep quality and sleepiness among shift nurses and health care workers. *International Journal of Occupational Safety and Ergonomics*, *21*(3), 312-317. https:/ / doi.org/10.1080/10803548.2015.1081770

Jakubowicz, D., Barnea, M., Wainstein, J., & Froy, O. (2013). High caloric intake at breakfast vs. dinner differentially influences weight loss of overweight and obese women. *Obesity*, *21*(12), 2504-2512. https:/ / doi.org/ 10.1002/oby.20460

Johnstone, A. (2014). Fasting for weight loss: An effective strategy or latest dieting trend? *International Journal of Obesity, 39*(5), 727–733. https://doi.org/10.1038/ijo.2014.214

Kasper, L. J., Alderson, R. M., & Hudec, K. L. (2012). Moderators of working memory deficits in children with attention-deficit/hyperactivity disorder (ADHD): A meta-analytic review. *Clinical Psychology Review*, *32*(7), 605-617. https://doi.org/10.1016/j.cpr.2012.07.001

Keller, M. C., Fredrickson, B. L., Ybarra, O., Côté, S., Johnson, K., Mikels, J., & Wager, T. (2005). A warm heart and a clear head: The contingent effects of weather on mood and cognition. *Psychological Science*, *16*(9), 724-731. https:/ / doi.org/ 10.1111/ j.1467-9280.2005.01602.x

Khedkar, P. H. (2020). Intermittent fasting

- The new lifestyle? *Acta Physiologica*, *229*(4), Article e13518. https://doi.org/10.1111/apha.13518

Lai, J. S., Hiles, S., Bisquera, A., Hure, A. J., McEvoy, M., &Attia, J. (2013). A systematic review and meta-analysis of dietary patterns and depression in community-dwelling adults. *The American Journal of Clinical Nutrition, 99(1),* 181–197. https://doi.org/10.3945/ajcn.113.069880

LeCheminant, J. D., Christenson, E., Bailey, B. W., & Tucker, L. A. (2013). Restricting night-time eating reduces daily energy intake in healthy young men: A short-term cross-over study. *British Journal of Nutrition*, *110*(11), 2108-2113. https://doi.org/10.1017/s0007114513001359

Leclerc, E., Trevizol, A. P., Grigolon, R. B., Subramaniapillai, M., McIntyre, R. S., Brietzke, E., & Mansur, R. B. (2020). The effect of caloric restriction on working memory in healthy non-obese adults. *CNS Spectrums*, *25*(1), 2-8. https://doi.org/10.1017/s1092852918001566

Lefevre, M., Redman, M., Heilbronn, L. K., Smith, J. V., Martin, C. K., Rood, J. C., Greenway, F. L., Williamson, D. A., Smith, S. R., & Ravussin, E. (2009). Caloric restriction alone and with exercise improves CVD risk in healthy non-obese individuals. *Atherosclerosis*, *203*(1), 206-213. https://doi.org/10.1016/j.atherosclerosis.2008.05.306

Liu, Z., Dai, X., Zhang, H., Shi, R., Hui, Y., Jin, X., Zhang, W., Wang, L., Wang, Q., Wang, D., Wang, J., Tan, X., Ren, B., Liu, X., Zhao, T., Wang, J., Pan, J., Yuan, T., Chu, C., ... & Liu, X. (2020). Gut microbiota mediates intermittent-fasting alleviation of diabetes-induced cognitive impairment. *Nature Communications*, *11*, Article 855. https://doi.org/10.1038/s41467-020-14676-4

Lo, J. C., Ong, J. L., Leong, R. L. F., Gooley, J. J., & Chee, M. W. L. (2016). Cognitive performance, sleepiness, and mood in partially sleep deprived adolescents: The need for sleep study. *Sleep, 39*(3), 687–698. https://doi.org/10.5665/sleep.5552

Longo, V. D., & Panda, S. (2016). Fasting, circadian rhythms, and time-restricted feeding in healthy lifespan. *Cell Metabolism, 23(6),* 1048–1059. https://doi.org/10.1016/j.cmet.2016.06.001

Manzanero, S., Erion, J. R., Santro, T., Steyn, F. J., Chen, C., Arumugam, T. V., & Stranahan, A. M. (2014). Intermittent fasting attenuates increases in neurogenesis after ischemia and reperfusion and improves recovery. *Journal of Cerebral Blood Flow & Metabolism*, *34*(5), 897-905. https://doi.org/10.1038/jcbfm.2014.36

Marosi, K., & Mattson, M. P. (2014). BDNF mediates adaptive brain and body responses to energetic challenges. *Trends in Endocrinology & Metabolism*, *25*(2), 89-98. https://doi.org/10.1016/j.tem.2013.10.006

Mattson, M. P. (2019). An Evolutionary Perspective on Why Food Overconsumption Impairs Cognition. *Trends in Cognitive Sciences, 23*(3), 200-212. https://doi.org/10.1016/j.tics.2019.01.003

McConnell, M. M., & Shore, D. I. (2011). Mixing measures: Testing an assumption of the Attention Network Test. *Attention, Perception, & Psychophysics*, *73*(4), 1096-1107. https://doi.org/10.3758/s13414-010-0085-3

Meule, A. (2020). The psychology of food cravings: The role of food deprivation. *Current Nutrition Reports*, *9*(3), 251–257. https://doi.org/10.1007/s13668-020-00326-0

Meule, A. (2023). *Assessment of Eating Behavior*. Hogrefe Publishing GmbH.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202. https://doi.org/10.1146/annurev.neuro.24.1.167

Mollayeva, T., Thurairajah, P., Burton, K., Mollayeva, S., Shapiro, C. M., & Colantonio, A. (2016). The Pittsburgh sleep quality index as a screening tool for sleep dysfunction in clinical and non-clinical samples: A systematic review and meta-analysis. *Sleep Medicine Reviews*, *25*, 52-73. https://doi.org/10.1016/j.sleep.2015.02.156

Moro, T., Tinsley, G., Bianco, A., Marcolin, G., Pacelli, Q. F., Battaglia, G., Palma, A., Gentil, P., Neri, M., & Paoli, A. (2016). Effects of eight weeks of time-restricted feeding (16/8) on basal metabolism, maximal strength, body composition, inflammation, and cardiovascular risk factors in resistance-trained males. *Journal of Translational Medicine*, *14*, Article 290. https://doi.org/10.1186/s12967-016-1044-0

Nebes, R. D., Buysse, D. J., Halligan, E. M., Houck, P. R., & Monk, T. H. (2009). Self-reported sleep quality predicts poor cognitive performance in healthy older adults. *The Journals of Gerontology: Series B*, *64*(2), 180-187. https://doi.org/10.1093/geronb/gbn037

Nugraha, B., Riat, A., Ghashang, S. K., Eljur-

nazi, L., & Gutenbrunner, C. (2020). A prospective clinical trial of prolonged fasting in healthy young males and females—Effect on fatigue, sleepiness, mood and body composition. *Nutrients*, *12*(8), Article 2281. https://doi.org/10.3390/nu12082281

Ooi, T. C., Meramat, A., Rajab, N. F., Shahar, S., Ismail, I. S., Azam, A. A., & Sharif, R. (2020). Intermittent fasting enhanced the cognitive function in older adults with mild cognitive impairment by inducing biochemical and metabolic changes: A 3-year progressive study. *Nutrients*, *12*(9), Article 2644. https://doi.org/10.3390/nu12092644

Patsalos, O., Keeler, J., Schmidt, U., Penninx, B. W., Young, A. H., & Himmerich, H. (2021). Diet, obesity, and depression: A systematic rview. *Journal of Personalized Medicine*, *11*(3), Article 176. https://-doi.org/10.3390/jpm11030176

Patterson, R. E., & Sears, D. D. (2017). Metabolic effects of intermittent fasting. *Annual Review of Nutrition*, *37*, 371-393. https://doi.org/10.1146/annurev-nutr-071816-064634

Patterson, R. E., Laughlin, G. A., Sears, D. D., LaCroix, A. Z., Marinac, C., Gallo, L. C., Hartman, S. J., Natajaran, L., Senger, C. M., Martinez, M. E., & Villaseñor, A. (2015). Intermittent fasting and human metabolic health. *Journal of the Academy of Nutrition and Dietetics*, *115*(8), 1203-1212. https://doi.org/10.1016/j.jand.2015.02.018

Psaltopoulou, T., Sergentanis, T. N., Panagiotakos, D. B., Sergentanis, I. N., Kosti, R., & Scarmeas, N. (2013). Mediterranean diet, stroke, cognitive impairment, and depression: A meta-analysis. *Annals of Neurology*, *74*(4), 580-591. https://doi.org/10.1002/ana.23944

Psychology Software Tools, Inc. [E-Prime 3.0]. (2016). Retrieved from https://support.pstnet.com/.

Rachid, H., Charaf, K., Hosbane, S., & Agoub, M. (2021). The benefits of Ramadan fasting on the cognitive function of medical students. *Journal of Nutrition, Fasting and Health*, *9*(2), 120-124. https://doi.org/10.22038/JNFH.2020.46756.1251

Rac-Lubashevsky, R., & Kessler, Y. (2016). Decomposing the n-back task: An individual differences study using the reference-back paradigm. *Neuropsychologia*, *90*, 190-199. https://doi.org/10.1016/j.neuropsychologia.2016.07.013

Radošević-Vidaček, B., Vidaček, S., & Kaliterna, L. (1990). The circadian rhythm parameters in mood variables. In E. Morgan (Ed.), *Chronobiology & Chronomedicine: Basic Research and Applications. Proceedings of the 4th Annual meeting of the European Society for Chronobiology* (pp. 286-294). Peter Lang.

Rossi, S., Lubin, A., Simon, G., Lanoë, C., Poirel, N., Cachia, A., & Houdé, O. (2013). Structural brain correlates of executive engagement in working memory: Children's inter-individual differences are reflected in the anterior insular cortex. *Neuropsychologia, 51*(7), 1145–1150. https://doi.org/10.1016/j.neuropsychologia.2013.03.011

Ruddick-Collins, L. C., Johnston, J. D., Morgan, P. J., & Johnstone, A. M. (2018). The Big Breakfast Study: Chrono-nutrition influence on energy expenditure and bodyweight. *Nutrition Bulletin*, *43*(2), 174-183. https://doi.org/10.1111/nbu.12323

Rynders, C. A., Thomas, E. A., Zaman, A., Pan, Z., Catenacci, V. A., & Melanson, E. L. (2019). Effectiveness of intermittent fasting and time-restricted feeding compared to continuous energy restriction for weight loss. *Nutrients*, *11*(10), Article 2442. https://doi.org/10.3390/nu11102442

Senderovich, H., Farahneh, O., & Waicus, S. (2023). The role of intermittent fasting and dieting on cognition in adult population: A systematic review of the randomized controlled trials. *Medical Principles and Practice*, *32*(2), 99–109. https://doi.org/10.1159/000530269

Short, M. A., & Louca, M. (2015). Sleep deprivation leads to mood deficits in healthy adolescents. *Sleep Medicine*, *16*(8), 987-993. https://doi.org/10.1016/j.sleep.2015.03.007

Solianik, R., & Sujeta, A. (2018). Two-day fasting evokes stress, but does not affect mood, brain activity, cognitive, psychomotor, and motor performance in overweight women. *Behavioural Brain Research, 338,* 166–172. https://doi.org/10.1016/j.bbr.2017.10.028

Teong, X. T., Hutchison, A. T., Liu, B., Wittert, G. A., Lange, K., Banks, S., & Heilbronn, L. K. (2021). Eight weeks of intermittent fasting versus calorie restriction does not alter eating behaviors, mood, sleep quality, quality of life and cognitive performance in women with overweight. *Nutrition Research, 92,* 32–39. https://doi.org/10.1016/j.nutres.2021.06.006

Welton, S., Minty, R., O'Driscoll, T., Willms, H.,

Poirier, D., Madden, S., & Kelly, L. (2020). Intermittent fasting and weight loss: Systematic review. *Canadian Family Physician Medecin de Famille Canadien*, *66*(2), 117–125.

Wesnes, K., & Pincock, C. (2002). Practice effects on cognitive tasks: a major problem? *The Lancet Neurology*, *1*(8), 473. https://doi.org/10.1016/s1474-4422(02)00236-3

Yamada, K., Mizuno, M., & Nabeshima, T. (2002). Role for brain-derived neurotrophic factor in learning and memory. *Life Sciences, 70*, 735-744. https://doi.org/10.1016/s0024-3205(01)01461-8

**Correction notice**
Incorrect Special Issue Labeling (Article erroneously excluded): This article was previously not labeled as part of a special issue due to an error. This has now been corrected.

# Cognitive or Emotional Improvement through Intermittent Fasting? Reflections on Hype and Reality

Stephan Schleim[1]

Intermittent fasting has received increasing scientific and public attention in recent years. The study by Batorek and Hromatko investigated whether time-restricted feeding, a form of intermittent fasting, improves cognitive performance and subjective-emotional well-being. This commentary discusses the most important results and relates them to previous studies on this topic. A major limitation of this new trial is its relatively short duration of only two months. I then link the idea of improving mental functions in healthy people to the discussion of cognitive or neuroenhancement. Finally, a current example of the communication of intermittent fasting in the media is discussed, which attracted public attention with a surprising message.

**Keywords** *hype, intermittent fasting, cognitive enhancement*

[1] Theory and History of Psychology, Heymans Institute for Psychological Research, Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, Netherlands

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
April 18, 2024
**Accepted**
May 24, 2024
**Published**
August 11, 2024

**Correspondence**
University of Groningen
s.schleim@rug.nl



People's desire to live as long and healthy a life as possible is reflected in scientific research today. Even in the distant past, people dreamed of a fountain of youth or even eternal life (Figure 1). Many universities now have research initiatives or even centers on such topics, for example under the heading of "healthy aging" (e.g., Center for Healthy Aging, n.d.; Leiden University, n.d.; University Medical Center Groningen, n.d.). Their focus is not only on lifespan in general, but in particular the extension of the mentally and physically *healthy* time as long as possible. Such efforts are also understandable in view of the fact that the average age is increasing in many prosperous societies. After all, various mental and physical limitations and illnesses occur more frequently in old age.

The new Special Issue of the *Journal of Trial and Error* now deals with "Scientific failure and uncertainty in the health domain." The present study "Cognitive functions, mood and sleep quality after 2 months of intermittent fasting" investigated the question of whether intermittent fasting improves cognitive performance and subjective well-being. Fasting (caloric reduction) and intermittent fasting in particular have received a lot of attention during the so-called "obesity epidemic" (Johnstone, 2015). This is reflected in the fact that not only the media and social media frequently cover the topic, but the number of scientific studies has also risen sharply in recent years (Figure 2). Intermittent fasting comes in many forms: For example, you can refrain from eating every other day, two days a week (the so-called 5:2 diet) or reduce the period within a day during which you eat. The latter is also known as "time-restricted feeding" (TRF) and was the ap-

**Companion Article**

Batorek & Hromatko (2024)

*Intermittent fasting: It makes one slimmer, but does it make one sharper?*

DOI: 10.36850/e71f-5cff

**Figure 1** The Fountain of Youth (1546) by Lucas Cranach the Elder (c. 1472-1553). The oil painting exhibited at the Gemäldegalerie Berlin depicts people's wish to remain youthful forever. License: public domain.

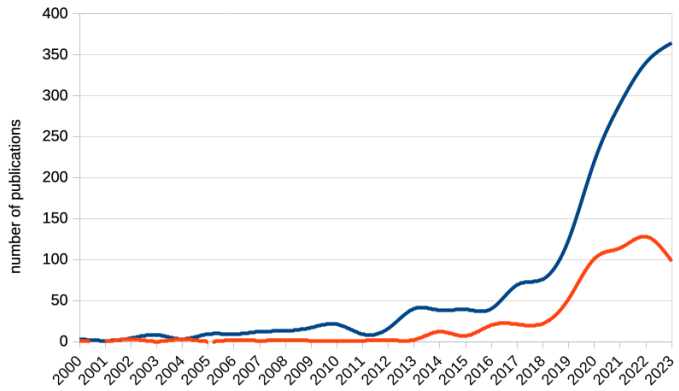



**Figure 2** Scientific publications on the topics of "intermittent fasting" (blue) and "time-restricted feeding" (red) in the *Web of Science*. The number of publications on the former has increased from less than one per week to more than one per day over the last ten years. From 2014 to 2023, there was almost a tenfold increase. Source: Web of Science (www.webofscience.com), Topic Search

proach taken in the new study. This can, for example, mean not eating for a period of 16 hours a day. In the following, I will describe and comment on the most important findings of the new study, discuss the broader research on cognitive performance enhancement and conclude with an outlook on the topic.

## The study

The researchers recruited 311 people to take part in the study via calls on social media. The researchers let them choose whether they wanted to stick to the diet (experimental group, TRF, $n$ = 231) or not change their eating habits (control group, $n$ = 80). Data on cognitive performance, sleep, and subjective well-being were collected at three points in time: shortly before starting the diet, after 1 month, and after 2 months, at the end of the study. In the two groups, 76 of the 231 (33%) and 29 of the 80 (36%) participants, respectively, remained. These people were on average 26 years old (range: 18 to 61) and had a slightly higher average body weight in the experimental group (70.6 vs. 65.8 kg) and a slightly higher body mass index (BMI: 24.1 vs. 22.6).

Cognitive tests were used to measure the subjects' attention and cognitive control. Memory span and working memory performance were also assessed. A questionnaire on sleep provided information on, among other things, the duration, interruption, and subjectively experienced quality of sleep. With regard to subjective well-being, the participants were asked to assess the extent to which predefined adjectives (e.g. about anxiety, energy, and happiness) applied to them. Although the results showed a statistically significant effect for the repetition of the various tests on cognitive performance, there was no significant difference between the groups. According to the researchers, the performance improved as a result of repeating the tests, thus a learning effect. Statistically, subjective well-being also improved significantly over time, but again there was no group effect. The authors suggest that this could be due to the fact that the weather improved over the course of the experiment (May to July 2021) and participants were on vacation. There was only a significant group difference in body weight: The BMI of the experimental group actually fell to an average of 23.3, bringing it closer to the value of the control group.

As a result, this study provides no evidence that intermittent fasting in the form of TRF increases cognitive performance or emotional well-being. The study authors already discuss some of the limitations of their study, in particular the lack of randomization. The reason for

this was to increase the intrinsic motivation of the participants for the diet. Since two-thirds of the participants in the experimental group stopped within the 2 months of the trial, it is reasonable to assume that the remaining 33% ($n$ = 76) were highly motivated, as there was no financial or other compensation for completing the experiment. In view of this fact, one might have expected at least a placebo or situation effect: Those who are more motivated perhaps exert themselves more in cognitive tests or at least answer the questions about subjective well-being more positively. However, the high motivation could of course have already been present at the first measurement, which would mean that it would no longer be visible in the comparison with the final result.

My biggest objection, though, concerns the short duration of the experimental intervention of only two months. The researchers cite two previous studies for their approach, which reported a positive effect of intermittent fasting on cognitive performance (Leclerc et al., 2020; Ooi et al., 2020). In these studies, however, the intervention lasted 24 and 36 months respectively, i.e. 12 to 18 times as long. In addition to that, Ooi and colleagues had examined a group of older people with symptoms of mild cognitive impairment (Ooi et al., 2020), while the participants in the new study were on average only 26 years old. Since possible changes in cognitive performance or subjective well-being must act via the brain, this raises the complex question of the extent to which a short diet of only two months can effectively improve brain activity. Additionally, the subjects in the experimental group were allowed a transitional phase of one week in which they were only supposed to fast for 14 hours (instead of the 16 hours that should ultimately be achieved) and were even encouraged to interrupt the diet one day a week and eat *ad libitem*. This further reduced the actual duration of the comparatively short intervention although it may have prevented even more dropouts.

A general question that I will address in more detail in the next section is that of the ecological validity of such tests. The neuropsychological tests frequently used in this type of research – as in Leclerc and colleagues (2020) – were originally developed to measure cognitive impairment in patients with psychological-psychiatric disorders and neurological diseases. It is still unclear what differences in these computer tests mean for the participants' everyday lives, which I described as a possible clinical and normative fallacy in earlier research (Schleim, 2014). The measured differences are also relative, even in the promising study by Leclerc and colleagues: The main statistically significant cognitive finding was that people in the experimental group made 23% fewer errors in a test of spatial working memory after 12 months and 32% fewer errors after 24 months; in the control group, they made 23% fewer errors after 12 months and 16% fewer errors after 24 months (Leclerc et al., 2020). That is a difference between the groups of 0% after 12 months and 16% after 24 months. Note that despite the randomization in this study, there was already a statistically significant different error rate between the two groups at the beginning of the dietary intervention (i.e., 18.9 vs. 24.6 points, $p < .05$), which casts doubt on the groups' cognitive homogeneity (Leclerc et al., 2020). This shows that these results are not only subject to fluctuations but must also be interpreted carefully.

The authors of the new study discuss the interesting idea that the diet could improve psychological functioning by improving sleep quality. This would also provide an answer to the aforementioned question of how brain function can be changed in the short term by fasting. Accordingly, a follow-up study with people with sleep disorders should be considered. However, the difficulties in their present study also reflect the complexity and heterogeneity of real life: People who would like to diet to improve their health and cognition will also be more or less motivated, undergo different seasons, go on vacation, and the like. This illustrates how both controlled experimentation and interventions in one's personal life can consist in trial and error – and yield ambiguous results. Even rigorous randomized controlled trials often cannot account for all possible confounds and need to simplify their boundary conditions, like the exclusion of patients with comorbidity in psychiatric research, while the patients in clinics often come with more than one mental problem (Hengartner, 2022).

## Cognitive enhancement

Since 2000, an academic discourse has gained

importance that deals with the improvement of cognitive performance (less frequently: mood, emotions) in healthy individuals. At first, this was referred to as "cognitive enhancement"; later, the term "neuroenhancement" was also used more commonly (Greely et al., 2008; Schleim & Quednow, 2018; Schleim, 2023). Following the "Decade of the Brain" (the 1990s), more thought was given to how people outside of therapeutic contexts could also benefit from neuroscience and neurotechnology. Probably due to the increasing use of psychopharmacological drugs, especially stimulants (such as methylphenidate/Ritalin or amphetamine), in the same period, such substances took a central role in the debate, even though methods of electrical or magnetic brain stimulation were discussed as well. Conventional or everyday activities such as physical exercise, sleep, and nutrition have also been discussed as ways of improving cognitive performance (Dresler et al., 2019). In this respect, the new study in the *Journal of Trial and Error,* too, can be classified in this area.

As I mentioned in the previous section, a diet would need to improve brain activity in order to improve mental function. The prefix "psycho" in "psychopharmacological drugs" indicates that these substances act directly on the nervous system. In this respect, psychoactive substances would be a more obvious attempt to influence cognitive performance and subjective well-being. However, this has proven to be much more difficult – in healthy and well-rested individuals – than was thought at the beginning of the neuroenhancement debate in the early 2000s (Schleim & Quednow, 2018; Schleim, 2023). For example, Roberts and colleagues (2020, pp. 20-21) conclude about the stimulants often prescribed for the treatment of attention disorders, when used by people without the diagnosis:

> *Methylphenidate has the strongest effects on cognition of the three stimulants observed. However, the positive effects are small to moderate, and limited to recall, inhibitory control and sustained attention. [...] D-amphetamine produces no improvements in cognition, and so can probably be ruled out of future investigation for safe, effective cognitive enhancement. The data with these stimulants is far from positive if we consider that effects are small and likely transient, in experiments that do not accurately reflect their actual use in the wider population.*

Even if this does not provide conclusive evidence, it relativizes the expectations of neuroenhancement – and indirectly also the cognitive improvement through diet: If the scientific results are sobering even with the substances that influence neurotransmitters like dopamine and noradrenaline in the brain, then improvement through food intake is likely to be even more difficult. Here Roberts and colleagues (2020), like me, point out the problem of ecological validity. A few more points or a couple of fewer errors in a neuropsychological test do not necessarily mean a more successful life (Schleim, 2014). In addition to the issue of ecological validity, the lack of long-term studies has also been pointed out for many years, which limits the scientific validity of many studies (Turner & Sahakian, 2006). This point also played a role in the discussion of the new study on intermittent fasting.

One of the rare exceptions among the already few studies with healthy test subjects asked experienced chess players to alternately play against a chess computer under the influence of placebo, caffeine, methylphenidate or modafinil in a double-blind study design (Franke et al., 2017). The increase in performance (games won) after taking the psychoactive substances was descriptively minimal and not statistically significant for all conditions. However, the study also had low power due to the small number of participants ($n$ = 39, all men). In a very competitive competition, though, where the performance of the competitors is close to each other, even small differences can make the difference between winning and losing. In this respect, it is interesting that the performance of the chess players was almost identical for caffeine and methylphenidate (both 4% more points than for placebo) and was even 2% higher under the influence of modafinil, a substance used to treat sleep disorders (Franke et al., 2017).

Therefore, in my opinion, such results show that we should avoid exaggerated expectations in research and discussion on the improvement of human beings. Studies on science communication, especially on genetic en-

gineering, have shown many years ago that scientists are under pressure to promise the applicability of their research ("translational pressure") and that scientists as well as their research institutes have an interest in getting media attention (Caulfield & Condit, 2012). I will discuss a current example in the field of diet research in the last section.

## Outlook: Hype, alarmism and reality

As already mentioned and shown above (Figure 2), the topics of fasting and interval fasting in particular have received a great deal of attention in recent years, both in the scientific community and in the general media. It is not currently foreseeable that this trend will change soon. While this commentary was being written, an interesting example occurred in the media that I would like to briefly discuss here. We could say that after the previous hype on intermittent fasting, we now meet an example of fasting-related alarmism. In any case, it illustrates the translational pressure to make it into the media.

On March 18, 2024, NBC News reported "Intermittent fasting linked to higher risk of cardiovascular death, research suggests" (Bendix, 2024), the *Washington Post* "The intermittent fasting trend may pose risks to your heart" (O'Connor, 2024) and a day later the Dutch NOS "American Heart Association warns against fasting, risk of premature death" (NOS Nieuws, 2024). The fact behind these far-reaching headlines is a poster by Chinese researchers at this year's American Heart Association Epidemiology and Prevention – Lifestyle and Cardiometabolic Health conference. These scientists are analyzing health data on behalf of the American Heart Association. On March 18, 2024, the Association published a press release about the poster entitled "8-hour time-restricted eating linked to a 91% higher risk of cardiovascular death", which attracted rapid media attention around the globe (American Heart Association, 2024). The results are preliminary and not yet peer-reviewed. The Dutch headline that the Heart Association itself warns against fasting can probably be regarded as a false report. Notably, the Association added the following caveat a day later: "As noted in all American Heart Association scientific meetings news releases, research abstracts are considered preliminary until published in a peer-reviewed scientific journal." A closer look at the poster in question shows that in the corresponding study on cardiovascular mortality, 423 of 11,831 people who ate for 12 to 16 hours a day, i.e. who did not follow a diet, died during the observation period, compared to 31 out of 414 people who ate for less than 8 hours a day, i.e. who applied intermittent fasting or TRF. From this, the researchers calculated a hazard ratio of 1.9, with a 95% confidence interval of 1.20 to 3.03. The 91% increased risk, which the press release and many media reports picked up on, is derived from this.

It would of course be dramatic if people improved their psychological functioning through a diet, but then died earlier because of a disease. However, as the study has not even been published yet, this news should be taken with at least a grain of salt. The whole issue becomes more complex when you consider that intermittent fasting is recommended to reduce obesity, which in turn is associated with cardiovascular disease itself (De Cabo & Mattson, 2019). This brings us to the contradictory scientific findings that fasting is associated with both more and less cardiovascular (and other) disease. However, as with all observational studies, the data presented on the above poster cannot be used to draw any firm causal inferences. The deaths of the 31 people could have been caused by factors that were not investigated in the study. In the discussion, however, muscle atrophy is brought into play as a possible mechanism for these deaths. Perhaps at least some of these people, who ate for less than 8 hours a day for many years, were simply undernourished?

The new study addresses a topic that is likely to remain relevant for the time to come. Intermittent fasting remains topical, as does the desire for an increase in cognitive performance and mood. To what extent nutrition plays a role not only in reducing or preventing symptoms of illness, but also in improving healthy living, has to be clarified by further research. It is already common wisdom that sufficient physical exercise, good sleep, and a balanced diet are also important for psychological well-being (Dresler et al., 2019; Schleim, 2023; Walsh, 2011). As an alternative to the term "neuroenhancement", I have proposed the term "instru-

mental substance use" to describe the use of psychoactive substances to improve life more generally (Schleim, 2020). Similarly, one could also speak of "instrumental fasting". However, there still seems to be a long way to go before concrete recommendations for everyday life can be derived from scientific studies.

## References


American Heart Association. (2024, March 18). *8-hour time-restricted eating linked to a 91% higher risk of cardiovascular death*. Newsroom. https://newsroom.heart.org/news/8-hour-time-restricted-eating-linked-to-a-91-higher-risk-of-cardiovascular-death/

Bendix, A. (2024, March 18). *Intermittent fasting linked to higher risk of cardiovascular death, research suggests*. NBC News. https://www.nbcnews.com/health/heart-health/intermittent-fasting-risk-cardiovascular-death-rcna143853

Caulfield, T., & Condit, C. (2012). Science and the sources of hype. *Public Health Genomics*, *15*(3-4), 209-217. https://doi.org/10.1159/000336533

Center for Healthy Aging (n.d.). *Center for Healthy Aging.* University of Copenhagen. https://healthyaging.ku.dk

De Cabo, R., & Mattson, M. P. (2019). Effects of intermittent fasting on health, aging, and disease. *New England Journal of Medicine*, *381*(26), 2541-2551. https://doi.org/10.1056/NEJMra1905136

Dresler, M., Sandberg, A., Bublitz, C., Ohla, K., Trenado, C., Mroczko-Wasowicz, A., Kühn, S., & Repantis, D. (2019). Hacking the brain: dimensions of cognitive enhancement. *ACS Chemical Neuroscience*, *10*(3), 1137-1148. https://doi.org/10.1021/acschemneuro.8b00571

Franke, A. G., Gränsmark, P., Agricola, A., Schühle, K., Rommel, T., Sebastian, A., Balló, H. E., Gorbulev, S., Gerdes, C., Frank, B., Ruckes, C., Tüscher, O., & Lieb, K. (2017). Methylphenidate, modafinil, and caffeine for cognitive enhancement in chess: A double-blind, randomised controlled trial. *European Neuropsychopharmacology*, *27*(3), 248-260. https://doi.org/10.1016/j.euroneuro.2017.01.006

Greely, H., Sahakian, B., Harris, J., Kessler, R. C., Gazzaniga, M., Campbell, P., & Farah, M. J. (2008). Towards responsible use of cognitive-enhancing drugs by the healthy. *Nature*, *456*(7223), 702-705. https://doi.org/10.1038/456702a

Hengartner, M. P. (2022). *Evidence-biased Antidepressant Prescription: Overmedicalisation, Flawed Research, and Conflicts of Interest*. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-82587-4

Johnstone, A. (2015). Fasting for weight loss: An effective strategy or latest dieting trend? *International Journal of Obesity*, *39*(5), 727-733. https://doi.org/10.1038/ijo.2014.214

Leclerc, E., Trevizol, A. P., Grigolon, R. B., Subramaniapillai, M., McIntyre, R. S., Brietzke, E., & Mansur, R. B. (2020). The effect of caloric restriction on working memory in healthy non-obese adults. *CNS Spectrums*, *25*(1), 2-8. https://doi.org/10.1017/S1092852918001566

Leiden University (n.d.). *Healthy Ageing.* Leiden University. https://www.universiteitleiden.nl/en/research-dossiers/taking-care-of-your-health/healthy-ageing

NOS Nieuws (2024, March 19). *Amerikaanse hartstichting waarschuwt voor vasten, kans op voortijdig overlijden.* NOS. https://nos.nl/artikel/2513387-amerikaanse-hartstichting-waarschuwt-voor-vasten-kans-op-voortijdig-overlijden

O'Connor, A. (2024, March 18). The intermittent fasting trend may pose risks to your heart. *The Washington Post.* https://www.washingtonpost.com/wellness/2024/03/18/intermittent-fasting-time-restricted-eating/

Ooi, T. C., Meramat, A., Rajab, N. F., Shahar, S., Ismail, I. S., Azam, A. A., & Sharif, R. (2020). Intermittent fasting enhanced the cognitive function in older adults with mild cognitive impairment by inducing biochemical and metabolic changes: A 3-year progressive study. *Nutrients*, *12*(9), Article 2644. https://doi.org/10.3390/nu12092644

Roberts, C. A., Jones, A., Sumnall, H., Gage, S. H., & Montgomery, C. (2020). How effective are pharmaceuticals for cognitive enhancement in healthy adults? A series of meta-analyses of cognitive performance during acute administration of modafinil, methylphenidate and D-amphetamine. *European Neuropsychopharmacology*, *38*, 40-62. https://doi.org/10.1016/j.euroneuro.2020.07.002

Schleim, S. (2014). Whose well-being? Common conceptions and misconceptions in the enhancement debate. *Frontiers in Systems Neu-*

*roscience*, 8, Article 148. https://doi.org/10.3389/fnsys.2014.00148

Schleim, S. (2020). Neuroenhancement as instrumental drug use: Putting the debate in a different frame. *Frontiers in Psychiatry*, *11*, Article 567497. https://doi.org/10.3389/fpsyt.2020.567497

Schleim, S. (2023). *Mental Health and Enhancement: Substance Use and Its Social Implications*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-32618-9

Schleim, S., & Quednow, B. B. (2018). How realistic are the scientific assumptions of the neuroenhancement debate? Assessing the pharmacological optimism and neuroenhancement prevalence hypotheses. *Frontiers in Pharmacology*, *9*, Article 315603. https://doi.org/10.3389/fphar.2018.00003

Turner, D. C., & Sahakian, B. J. (2006). Neuroethics of Cognitive Enhancement. *BioSocieties*, *1*(1), 113-123.

University Medical Center Groningen (n.d.). *Healthy Ageing*. UMCG Research. https://umcgresearch.org/w/healthy-ageing

Walsh, R. (2011). Lifestyle and mental health. *American Psychologist*, *66*(7), 579-592. https://doi.org/10.1037/a0021769

**Correction notice**
Incorrect Special Issue Labeling (Article erroneously excluded): This article was previously not labeled as part of a special issue due to an error. This has now been corrected.

# Smile, You're on Camera: Investigating the Relationship between Selfie Smiles and Distress

Monika Lind[1,2], Michelle Byrne[1,3], Sean Devine[4,†], Nicholas Allen[1]

Background: This study examined the relationship between (1) participant smiling in daily "selfie" videos and (2) self-reported distress. Given the extensive use of digital devices for sharing expressions of non-verbal behavior, and some speculation that these expressions may reveal psychological states—including emotional distress—we wanted to understand whether facial expression in these TikTok-like videos were correlated with standardized measures of psychological distress. Based on the work of Paul Ekman and others, which posits that facial expressions are universal reflections of people's inner states, we predicted that smiling would be inversely related to psychological distress. Method: Twenty-four undergraduate students, aged 18+ years *(M = 18.35, SD = 2.75)*, were prompted to record a two-minute selfie video each evening during two weeks of data collection (i.e., 14 total days). They were instructed to describe various aspects of their day. They also completed self-report questionnaires at the end of each assessment week, including the Depression Anxiety Stress Scale (DASS), Perceived Stress Scale (PSS), and the Pittsburgh Sleep Quality Index (PSQI). Results: A counterintuitive effect was observed whereby smiling intensity during selfie videos was positively correlated with individual differences in anxiety, depression, and stress. Discussion: This study challenges the common view that facial expressions necessarily reflect our inner emotions. It provides preliminary evidence that a mobile sensing app that captures selfies—along with other naturalistic data—may help elucidate the relationship between facial expressions and emotions.

**Keywords** *mobile sensing, video, selfie, facial expression, smile*

[1]Department of Psychology, University of Oregon, Eugene, United States
[2]School of Social Ecology, University of California, Irvine, United States
[3]Turner Institute for Brain and Mental Health, Monash University, Melbourne Australia
[4]Department of Psychology, McGill University, Montreal, Canada

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
March 3, 2022
**Accepted**
March 6, 2024
**Published**
May 24, 2024

**Correspondence**
University of California
m.lind@uci.edu



This is a small study with counterintuitive findings that bear on our understanding of facial expressions and our use of selfie data. This study lands on the growing stack of publications that support the view that facial expressions are something other than an accurate readout of a person's internal state (Barrett et al., 2019). Furthermore, this study shows that a special aspect of our lab's mobile sensing app, the ability to collect *TikTok*-like "selfie" videos, may provide meaningful insights into psychological states.

This study draws its data from the Effortless Assessment of Stressful Experiences project, a small but mighty pilot. The project aimed to find out whether a mobile sensing app, the Effortless Assessment Research System (EARS; Figure 1), could produce meaningful data about distress without EARS itself driving research participants crazy. This study builds on two existing publications from the project:

[†]During the peer review process Sean Devine peer reviewed the article and signed his review, in which he pointed out that the pre-planned analysis was not the most efficient use of the data, given the data's nested structure. Devine provided multiple alternatives for how to model the data. The authors subsequently contacted Devine to join as an author. After this occured and changes were made to the article, the article again went through independent peer review.

## Take-home message

This is a small study with counterintuitive findings. It further loosens the grip of the common view that our facial expressions necessarily reflect our inner emotional states, and it provides preliminary evidence that a mobile sensing app that captures selfies can help elucidate the relationship between facial expressions and emotions.

a paper introducing EARS, including information about its excellent acceptability (Lind et al., 2018), and a paper on the associations between EARS data (typed text), self-reported stress, and biological markers of inflammation (Byrne et al., 2021).

This study focuses on the relationship between (1) participant smiling in selfie videos captured by EARS and (2) self-reported distress. The field of facial expression research is rich with nuanced research and theories that defy distillation...but that won't stop me from trying! Paul Ekman advanced the expressed emotion framework for facial expressions, which posited that facial expressions are universal reflections of people's inner states (Ekman, 1993; Ekman & Friesen, 1971). In a classic academic plot twist, Alan Fridlund, Dr. Ekman's junior collaborator, contradicted his thesis (Figure 2). Dr. Fridlund took a behavioral ecology view of facial expressions that posited that facial expressions primarily serve a person's communication goals and therefo
align with their inner emotic

This debate continues to animate facial expression research specifically and emotion research more broadly (Ekman, 2016). Fridlund's behavioral ecology view of facial expressions leads a robust field of research advancing motivation-communication views, while Lisa Feldman Barrett has led the development of constructivist emotion theories, which characterize emotion and facial expressions as comprised of orthogonal dimensions of valence and arousal (Barrett, 2006; Mayo & Heilig, 2019). A growing wave of evidence undermines Ekman's framework, including studies conducted in the same settings where Ekman first found support for his views (i.e., in small, indigenous societies; Crivelli et al., 2017; Gendron et al., 2018). This wave has a sturdy bulwark to overcome: In a survey of about 150 top researchers of emotion, 80% of respondents endorsed an Ekman-aligned view that facial expressions provide universally interpretable read-outs of emotions (Ekman, 2016).

Current research on selfies tends to focus on publicly available images, e.g., images on Instagram. Deeb-Swihart and colleagues (2017) used computer vision and network analysis techniques on 2.5 million selfies from Instagram, the largest empirical analysis of selfies at the time. They categorized the selfies into typologies of identity statements and found that the categories present in online selfies reflect the same categories present in real-life identity statements (e.g., wealth, health,

**Figure 1** EARS app icon.

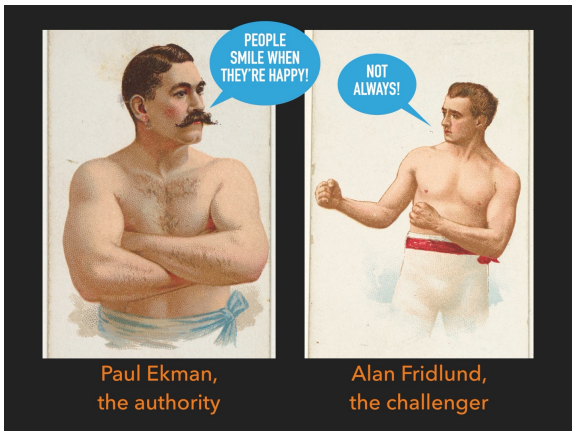



**Figure 2** Dr. Ekman and Dr. Fridlund; images from the public domain.

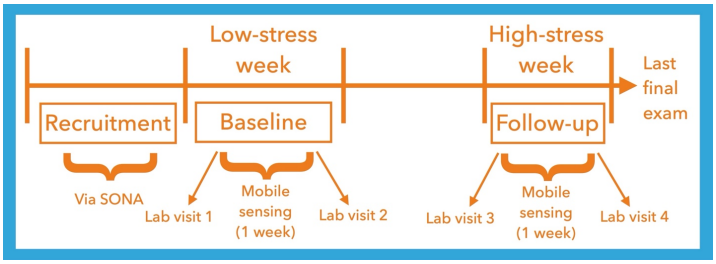



**Figure 3** Timeline of the EASE study.

and physical attractiveness). This focus on the identity-related content of selfies typifies current selfie research. Psychological studies of facial expressions in selfie images and videos tend to focus on the degree to which human judgment and machine learning approaches can accurately perceive the personality traits of the selfie subjects (Kachur et al., 2020; Qiu et al., 2015). Meanwhile, medical studies of facial expressions in selfies have produced promising findings related to the detection of illnesses that affect the expressiveness of the face, e.g., Parkinson's Disease (Grammatikopoulou et al., 2019). We believe this is the first study to focus on the relationship between facial expressions in selfie videos and psychological distress.

Valuing the simplicity of the expressed emotion framework and recognizing it as the default view in the field, we decided to align our hypotheses with Ekman. We hypothesized that within-subject increases in self-reported distress would be associated with decreases in smiling intensity.

## Method

### Protocol

We used an academic stress paradigm to test this hypothesis (Figure 3). The academic stress paradigm takes advantage of the varying degrees of stress built into each academic term (Zunhammer et al., 2013). In our case, the academic stress paradigm let us gather mobile sensing and questionnaire data twice: once during a low-stress week (nothing major due) and once during a high-stress week (finals week). We determined the specific data-gathering schedule for each participant based on the participant's report of their major assignment deadlines and their last final exam. (If you, dear reader, are on the quarter system and you're thinking, "There are no low-stress weeks!" *ahem*, hang onto that thought.)

### Participants

Undergraduates at a large public university in the American Northwest enrolled in the study via the Psychology and Linguistics human subjects pool during the 2016-17 academic year. Inclusion criteria were using an Android phone and not having an immunological medical condition (which was related to a study aim not reported in this paper). All participants used Android phones because this classic 2016 vintage of EARS only worked on Android. During the study period, reports indicated that Android users comprised at least half of the American smart phone market (Comscore, 2016; Siegal, 2017). A contemporaneous study showed that there was little empirical support for personality differences between Android and iOS users at that time (Götz et al., 2017). Additional contemporaneous research on the characteristics of iPhone versus Android users is lacking. As such, a pilot study with only Android users is reasonable. Participants received $50 compensation. The university's Institutional Review Board approved the study (protocol number 07212016.019), and participants provided informed consent during the first lab visit.

Participants were 25 undergraduate students, aged 18+ years ($M$ = 18.35, $SD$ = 2.75). Twenty-four participants completed the whole study. Twelve participants (50%) were cisgender female, and 12 (50%) were cisgender male; 12% were Asian, 64% Caucasian, 12% Hispanic, and 12% Multiracial. The yearly income of participants' parents ranged from lower to upper-middle class ($M$ = $88,625.00, $SD$ = 62,009.69).

### Measures

To collect facial expression data, EARS prompted each participant to record a two-minute selfie video each evening during data collection (14 total days). In each selfie video, we instructed participants to state their name, date, time, and the weather. We then prompted them to describe one thing that happened that day that was positive ("What was the best thing that happened today, and

can you describe it?"), and one thing that was negative ("What was the most difficult thing that happened today, and can you describe it?").

As part of the battery of self-report questionnaires at the end of each assessment week, participants completed the Depression Anxiety Stress Scales (DASS), Perceived Stress Scale (PSS), and the Pittsburgh Sleep Quality Index (PSQI). The DASS is a 42-item questionnaire designed to measure the three related negative emotional states of depression, anxiety, and tension/stress (Lovibond & Lovibond, 1995). This version of the DASS has high internal consistency across clinical and community samples (Antony et al., 1998). The PSS is a 14-item questionnaire designed to measure perceived stress (Cohen et al., 1983). This version of the PSS has Cronbach's alpha values consistently over 0.70 across multiple cultures and countries (Lee, 2012). The PSQI measures subjective sleep quality and quantity (Buysse et al., 1989). For the DASS, PSS, and PSQI, we instructed participants to respond based on their experiences over the last week.

At baseline, participants completed the computerized version of the Stress and Adversity Inventory (STRAIN), an NIMH/RDoC-recommended instrument for measuring cumulative exposure to life stress (see https: / / www.strainsetup.com/ ; Slavich & Shields, 2018). The STRAIN uses intelligent logic to shape the interview based on the participant's responses.

### Data Analysis

The 24 participants generated 325 selfie videos. We used OpenFace, a "facial behavior analysis toolkit," to conduct automated analysis of participants' smiles. OpenFace output includes facial landmark detection, facial landmark and head pose tracking, facial action unit recognition, gaze tracking, and facial feature extraction (Baltrusaitis et al., 2018). Action units (AUs) refer to small movements of the face, like action unit 6 ("the cheek raiser") or action unit 9 ("the nose wrinkler"). We focused on facial action unit 12 (AU12), commonly referred to as the "lip corner puller," which captures smiling. Smiling warrants special attention for two reasons: One, the smile is the most common facial expression in the datasets used to develop facial expression analysis software, so it is the facial expression that OpenFace detects most reliably; and two, smiling is meaningful around the world and relates to important life outcomes (Godoy et al., 2005). OpenFace produces both a categorical present/absent measure of action units and a continuous intensity measure of action units. We focused on the intensity measure (range: 0-5) to capture nuance in smiling behavior beyond just presence and absence. We chose not to check the content of the selfie videos because we wanted to determine whether fully automated facial expression analysis (i.e., without a human in the loop), was a viable approach for future, large-scale studies.

To test our hypothesis that smiling intensity would decrease as distress increased, we planned to leverage our within-subjects design to test whether change in smiling during selfie videos between the low-stress week (Time 1) and the high-stress week (Time 2) would predict change in self-reported symptoms over the same assessment periods. (Because we did not have daily questionnaire scores, we were unable to test the association of daily smiling with daily distress.) Our plan entailed calculating Time 1 and Time 2 averages for smiling and questionnaires, regressing Time 2 scores on Time 1 scores for smiling and questionnaires, then regressing the residuals of the questionnaire scores on the residuals of the smile intensity scores. Due to the small number of confirmatory hypotheses, we did not correct for multiple comparisons.

## Results

### Hypothesis Testing

First, we checked to see if the academic stress paradigm had the desired effect: Did participants report feeling more distressed during the high-stress week? Yes, for the most part! Second, we checked smiling behavior: Did it change in the predicted direction as well? Yes, participants smiled less—numerically but not statistically significantly, *hmmm*—in the high-stress week! These patterns (Table 1) presaged well for our hypothesis that increases in distress would be associated with decreases in smiling, and we were feeling pretty clever *if we do say so ourselves*. When we tested our

hypothesis by regressing the residuals of the symptom scores on the residuals of the smile intensity scores, however, non-significant results rudely contradicted our cleverness. Anxiety residuals regressed on smiling residuals yielded: $F(1,22) = 1.04$, $p$ = NS, Adjusted $R^2$ = 0.002.

We made a final attempt at testing our hypothesis during the peer review process. We received a thoughtful, helpful, *signed* review that pointed out that our pre-planned analysis was not the most efficient use of our data, given our data's nested structure. The reviewer provided multiple alternatives for how to model our data. With the blessing of the editor, we contacted the reviewer and invited him to join the paper. The reviewer —now the third author— conducted a multilevel multiple regression with smiling intensity per video nested within participant as the outcome and change scores of self-reported distress and sleep crossed with week (low-stress vs high-stress) as the predictors (Table 2). One way to articulate this model in lay terms is, "Does smiling decrease in high-stress weeks for a person who experienced that week as more distressing?"

Look at those p-values! They're huge! (Relatively speaking, of course— we all know p-values are not measures of magnitude.) While the direction of the simple fixed effects aligned with our hypothesis that increased distress would predict decreased smiling, we once again fell far short of statistical significance.

Our null findings forced us to reckon with an inconvenient truth: The academic stress paradigm may not have worked very well. It's true that we saw changes in the hypothesized directions, but these changes were small and only statistically significant in the case of self-reported anxiety ($t(23) = -2.14$, $p < .05$).

## Exploratory Analyses

From this point forward, we conducted exploratory data analysis. Given the failure of the academic stress paradigm, we decided to treat week one and week two data as multiple measurements of the same constructs. We collapsed across the weeks, calculating averages for each participant for each variable of interest (smiling, stress, anxiety, depression, and sleep). We calculated differences in smile intensity by gender and racial and ethnic identity, but we were underpowered to test the significance of these differences (Tables 3 and 4).

We moved on to exploring the relationships among the key variables. There we were, plotting bivariate associations in the hopes of resurrecting our hypothesis, when all of a sudden:

That looks like higher anxiety is associated with... more smiling? A perusal of the correlation table fleshed out the story (Table 5).

We set aside significance testing because of the exploratory nature of these analyses and focused instead on effect sizes. Following Cohen's (1992) guidance, we had three medium and three small effect sizes. In our data, smiling intensity had a medium, positive association

**Table 1** Means, standard deviations, and ranges of smiling intensity, DASS subscales (Depression, Anxiety, Stress), perceived stress (PSS), and self-reported sleep quantity in hours (PSQI) by week, plus the change in the means across weeks.

| | Low-Stress Week | | | High-Stress Week | | | Change |
|---|---|---|---|---|---|---|---|
| | Mean | *SD* | Range | Mean | *SD* | Range | |
| Smiling | 0.60 | 0.44 | 0.02-1.60 | 0.55 | 0.43 | 0.02-1.65 | -0.05 |
| DASS-Dep | 7.42 | 6.88 | 0.00-23.00 | 6.46 | 5.51 | 0.00-23.00 | -0.96 |
| DASS-Anx | 5.46 | 5.00 | 0.00-17.00 | 7.67 | 5.54 | 0.00-21.00 | +2.21 |
| DASS-Str | 11.00 | 8.24 | 0.00-26.00 | 12.84 | 9.06 | 1.00-33.00 | +1.84 |
| PSS | 24.57 | 8.50 | 6.00-37.00 | 25.04 | 10.05 | 4.00-40.00 | +0.47 |
| Sleep | 6.78 | 1.30 | 4.00-9.00 | 6.98 | 1.52 | 2.00-9.00 | +0.20 |

**Table 2** Multilevel model regressing smiling inte within participant on standardized change scores scales and sleep crossed with week.

| Predictor | Estimate | 95 |
|---|---|---|
| (Intercept) | -0.02 | -0.36 – |
| Week | -0.08 | -0.27 – |
| Anxiety change | -0.08 | -0.71 – |
| Depression change | -0.23 | -0.69 – |
| Stress change | -0.01 | -0.55 – |
| Sleep change | 0.18 | -0.16 – |
| Week x Anxiety change | 0.08 | -0.26 – |
| Week x Depression change | 0.01 | -0.24 – |
| Week x Stress change | 0.05 | -0.24 – |
| Week x Sleep change | -0.08 | -0.28 – |

Marginal $R^2$: 0.117
Conditional $R^2$: 0.747

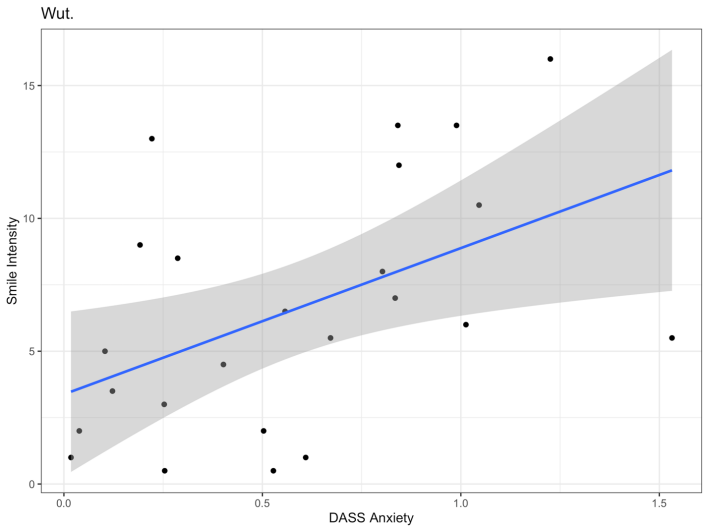



**Figure 4** Anxiety and smiling intensity increase together.

**Table 3** Mean and standard deviation of smiling intensity by gender.

| Gender | *N* | Smiling Intensity |
|---|---|---|
| Cisgender Female | 12 | 0.68 (*SD* = 0.46) |
| Cisgender Male | 12 | 0.47 (*SD* = 0.34) |

**Table 4** Mean and standard deviation of smiling intensity by racial and ethnic identity.

| Race/Ethnicity | *N* | Smiling Intensity |
|---|---|---|
| Asian | 3 | 0.61 (*SD* = 0.38) |
| Hispanic | 3 | 0.21 (*SD* = 0.17) |
| Multiracial | 3 | 0.45 (*SD* = 0.31) |
| White | 15 | 0.67 (*SD* = 0.44) |

with anxiety, depression, and stress; a small, positive association with perceived stress and lifetime stress; and a small, negative association with self-reported sleep quantity. In other words, smiling intensity, anxiety, depression, and stress tended to increase together.

Further investigation while accounting for multiple measurements of smiling behavior mirrored this pattern. We conducted a series of simple multilevel regressions with smiling intensity per video nested within participant as the outcome. Average depression, anxiety, stress, and sleep each took its turn as the predictor (Table 6).

While these findings may be counterintuitive, their essence—that we may smile through stress—already pervades American popular culture:


**Figure 5** The "This is fine" dog, from the first two panels of KC Green's comic, "On Fire," from the series, "Gunshow" (Green, 2013); used with permission.

Do you know for sure why the dog is smiling despite the flames and smoke? We don't. The same is true of the results of this study: We know we saw an unexpected, compelling pattern in the data, and we don't know why. We'd like to offer some ways to think about these results.

**Table 5** Pearson correlations and 95% confidence intervals for smiling intensity, the three DASS subscales, perceived stress (PSS), lifetime stress exposure (STRAIN), and self-reported sleep quantity in hours.

| | Smiling | DASS-Dep | DASS-Anx | DASS-Str | PSS | STRAIN | Sleep |
|---|---|---|---|---|---|---|---|
| **Smiling** | 1.00 | 0.38[-.03, .68] | 0.48[.10, .74] | 0.37[-.03, .68] | 0.29[-.13, .62] | 0.18[-.26, .56] | -0.24[-.58, .19] |
| **DASS-Dep** | | 1.00 | 0.52 | 0.43 | 0.45 | 0.02 | -0.20 |
| **DASS-Anx** | | | 1.00 | 0.79 | 0.66 | 0.37 | -0.27 |
| **DASS-Str** | | | | 1.00 | 0.72 | 0.36 | -0.27 |
| **PSS** | | | | | 1.00 | 0.27 | -0.46 |
| **STRAIN** | | | | | | 1.00 | -0.47 |
| **Sleep** | | | | | | | 1.00 |

## Discussion

Our original hypothesis rested on the expressed emotion framework advanced by Paul Ekman, which conflicted with the behavioral ecology view of Alan Fridlund. This study, with its finding that smiling increased alongside stress, anxiety, and depression, may lend more support to Dr. Fridlund's view that facial expressions primarily serve a person's communication goals. For example, participants experiencing more anxiety might be more worried about wanting the implied audience (e.g., researchers) to like them and their videos, so they might smile more. However, we cannot say conclusively that these findings support Dr. Fridlund's view – only that in a dichotomous scenario where Dr. Ekman and Dr. Fridlund oppose each other (such as, say, a boxing match), these findings align more with Dr. Fridlund.

The complexity of the smile as a behavioral expression could also help us think about this study's results. In facial expression research, we often use action units to describe facial configurations. To capture smiling, we measured only one facial action unit: Action Unit 12, the "lip corner puller". Studies have found, however, that different types of smiles emerge when AU12 is combined with other action units. Dr. Ekman himself argued that there are "felt, false, and miserable smiles," and Magdalena Rychlowska and colleagues provided a similar, more data-driven structure of three different categories of smiles: reward, affiliative, and dominance (Ekman & Friesen, 1982; Rychlowska et al., 2017). Our simplistic index of smiling did not allow us to capture this complexity. If we were able to break the smiles down by type, we might find that one type of smile drives the association between smiling and distress.

Participants tolerated the EARS tool well, as evidenced by only one participant dropping out, zero participants needing to use the provided battery packs, and zero participants reporting interference with usual phone usage. Still, the selfie video task itself deserves some scrutiny. Each day, participants pointed their front-facing phone cameras at themselves and recorded for 30+ seconds while, in true selfie fashion, seeing themselves reflected back on their phone screens. What happens to us when we look at ourselves? How does the little Zoom box of your face affect your facial expressions during meetings? What if you're already feeling a bit stressed? Duval and Wicklund's objective self-awareness theory states that when people are made consciously aware of themselves, they compare themselves to their own

**Table 6** Four simple multilevel models regressing nested smiling behavior on standardized, averaged self-report measures (three DASS subscales and PSQI sleep quantity).

| Predictor | Estimate | 95% CI |
|---|---|---|
| Depression | 0.34 | 0.02 – 0.67 |
| Anxiety | 0.44 | 0.13 – 0.75 |
| Stress | 0.35 | 0.01 – 0.68 |
| Sleep | -0.23 | -0.54 – 0.08 |

standards (Duval & Wicklund, 1972). Perhaps the reflection of participants' own faces increased their self-awareness causing them to compare themselves to their standards and smile more. This potential demand characteristic might have been exacerbated by participant anxiety and stress.

Finally, discussion of this study must take stock of the ways that it aligns with and falls short of Barrett and colleagues' recommendations in their 2019 instant classic review (Barrett et al., 2019). Our study aligns with Barrett and colleagues' calls to study facial movements in real life, combine classical psychology methods (i.e., self-report) with machine learning (i.e., automated facial analysis via OpenFace), use novel methods (i.e., mobile sensing and OpenFace), and design studies that allow findings outside of the *common view* that a certain set of facial expressions reflect our inner emotional states. On the other hand, our study falls short of Barrett and colleagues' calls to conduct larger scale studies that produce rich data across contexts, identify neural mechanisms of facial expressions, map the dynamics of facial actions, and avoid perpetuating past errors in emotion research (i.e., choosing by default a *common view*-inspired hypothesis).

### Limitations

Various aspects of the study design limit these findings. First, the academic stress paradigm may not have worked very well. We think the minimal contrast between the low-stress week and the high-stress week has to do with the dreaded quarter system. Our intrepid research assistants got an inkling of this issue during data collection when they struggled to schedule initial lab visits because it was so hard to identify a week during the term when participants had nothing major due. (For an accounting of this and all of the first author's other errors, see Box 1 toward the end of this article.) It's also possible that increasing familiarity with the selfie video task, i.e., practice effects, may have obscured differences in naturalistic smiling behavior between weeks. It's unlikely that seasonal effects bear on our results as the study was conducted in both fall and spring terms.

In addition, at three key points in the protocol, we missed opportunities to gather helpful data. First –and most embarrassingly– we failed to measure positive affect in a study focused on smiling, which makes it impossible to say anything about how smiling might have related to positive emotions. Second, we failed to use ecological momentary assessment to ask participants about their moods when they recorded each selfie video. Third, we failed to ask participants during debriefing about who they imagined the audience to be for their selfie videos. The selfie videos carried with them two additional limitations. We instructed participants to talk throughout their videos, which is likely to have introduced extra noise to the automated facial analysis. We also opted not to hand code or check OpenFace's analysis of the videos, which means that our analysis did not benefit from a human in the loop.

Finally, these exploratory findings are based on a small, predominately white sample of college students. A pilot study of 24 mostly monochromatic undergrads cannot support strong conclusions of any kind. Rather than carry on for another paragraph about *future directions*, we'll identify just one: replication.

### Unforeseen Evils

Having ventured into perilous digital mental health territory, we must learn from the valiant scientists who have gone before us and specify what our findings *do not* mean. This study provides no evidence that the selfie videos in any way cause distress. This is a correlational study that suggests that distress and smiling, in this particular context, tend to increase together. Furthermore, these findings should be

## Original purpose

Mobile sensing is the collection of naturalistic behavioral data through digital means, including methods that use an app installed on a participant's phone with their informed consent. We wanted to find out whether selfie videos from a mobile sensing app could produce meaningful data about distress. We focused on the relationship between (1) participant smiling in selfie videos captured via mobile sensing and (2) self-reported distress. We hypothesized that increases in distress would be associated with decreases in smiling.

interpreted in the context of revelations about the biases baked into many automated facial analysis tools. The training datasets that underpin OpenFace vary in both their intentional inclusion of diverse faces and their disclosure of the dataset's demographics. As we were underpowered to test for significant differences in smiling intensity between participants based on racial or ethnic identity in this study (Table 4), it is appropriate to proceed with caution with the application of automated facial analysis to racially and ethnically diverse samples.

## Box 1: An Incomplete Catalogue of the First Author's Errors

- We made a mess out of our video file naming conventions over multiple updates of the EARS app. Think underscores in some, spaces in others. Standardize your file naming conventions, people!
- We ran the academic stress paradigm on the quarter system. As anyone on the quarter system knows, it can feel like a flat-out sprint from week one till finals. We think this undermined the academic stress paradigm, and we would recommend running it only on the semester system.
- When it became clear that the academic stress paradigm didn't work well, I decided to "collapse the data across weeks." Because I was a wee baby grad student, my first attempt at this was to take my 24 participants with 2 sets of observations and lengthen the dataset to 48 total observations... without accounting for repeated measurements. This inflated my effect sizes a bit! I corrected this mistake by averaging the repeated measurements.
- We first ran the automated facial expression analysis using OpenFace in 2016. Various factors delayed my drafting of this paper (no, *you're* anxious about academic writing!) such that enough time elapsed for a new version of OpenFace to come out. The developer reported significant improvements with the new version, so we had to run the analysis again.
- When I was first scoring the DASS-42, I used the DASS-21 scoring guide, which... dramatically underestimated symptoms.

## Conclusion

This small but mighty pilot study loosens the grip of the *common view* that our facial expressions reflect our inner emotional states, and it provides preliminary evidence that a mobile sensing app that captures selfies along with other naturalistic data—may help elucidate the relationship between facial expressions and emotions. It doesn't prove anything grandiose. Rather, we hope that our presentation of this study, warts and all, provides useful lessons gained incrementally through the laudable scientific process of trial and error.

## Acknowledgements

The authors wish to thank the participants in this study for their time and effort. In addition, the authors acknowledge the valuable contributions of: Jeff Cohn, LP Morency, Jeff Girard, Laszlo Jeni, Wen-Sheng Chu, and Nicki Siverling, for teaching the first author about automated facial analysis; Tadas Baltrusaitas, for analyzing the selfie videos using OpenFace; Lauren Kahn, for assisting with data analysis and providing feedback on an early draft of this manuscript; Sanjay Srivastava, for providing feedback on an early draft of this manuscript; and Elizabeth McNeilly, for assisting with data analysis.

## Funding Acknowledgements

This study was funded by the Stress Measurement Network via a grant from the National Institute on Aging (R24AG048024).

## Conflicts of Interest

The authors declared the following potential conflicts of interest with respect to the research, authorship, or publication of this article: Smile, you're on camera: Investigating the relationship between selfie smiles and distress. Monika Lind, Michelle Byrne, and Nicholas Allen hold equity interests in Ksana Health Inc., a company that has the sole commercial license for certain versions of the Effortless Assessment Research System (EARS) mobile phone application and some related EARS tools. The authors have nothing else to disclose.

## References


Antony, M. M., Bieling, P. J., Cox, B. J., Enns, M. W., & Swinson, R. P. (1998). Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychological Assessment*, *10*(2), 176–181. https://doi.org/10.1037/1040-3590.10.2.176

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 59–66. https://doi.org/10.1109/FG.2018.00019

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, *1*(1), 28–58. https://doi.org/10.1111/j.1745-6916.2006.00003.x

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research*, *28*(2), 193–213. https://doi.org/10.1016/0165-1781(89)90047-4

Byrne, M. L., Lind, M. N., Horn, S. R., Mills, K. L., Nelson, B. W., Barnes, M. L., Slavich, G. M., & Allen, N. B. (2021). Using mobile sensing data to assess stress: Associations with perceived and lifetime stress, mental health, sleep, and inflammation. *DIGITAL HEALTH*, *7*. https://doi.org/10.1177/20552076211037227

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.155

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A Global Measure of Perceived Stress. *Journal of Health and Social Behavior*, *24*(4), 385–396. https://doi.org/10.2307/2136404

Comscore. (2016, March 3). *Comscore Reports January 2016 U.S. Smartphone Subscriber Market Share*. https://www.comscore.com/Insights/Rankings/comScore-Reports-January-2016-US-Smartphone-Subscriber-Market-Share

Crivelli, C., Russell, J. A., Jarillo, S., & Fernández-Dols, J.-M. (2017). Recognizing spontaneous facial expressions of emotion in a small-scale society of Papua New Guinea. *Emotion*, *17*, 337–347. https://doi.org/10.1037/emo0000236

Deeb-Swihart, J., Polack, C., Gilbert, E., & Essa, I. (2017). Selfie-presentation in everyday life: A large-scale characterization of selfie contexts on Instagram. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 42–51. https://doi.org/10.1609/icwsm.v11i1.14896

Duval, S., & Wicklund, R. A. (1972). *A theory of objective self awareness* (pp. x, 238). Academic Press.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384–392. https://doi.org/10.1037/0003-066X.48.4.384

Ekman, P. (2016). What scientists who study emotion agree about. *Perspectives on Psychological Science*, *11*(1), 31–34. https://doi.org/10.1177/1745691615596992

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129. https://doi.org/10.1037/h0030377

Ekman, P., & Friesen, W. V. (1982). Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, *6*(4), 238–252. https://doi.org/10.1007/BF00987191

Fridlund, A. J. (1994). *Human Facial Expression: An Evolutionary View*. Academic Press.

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality reconsidered: Diversity in making meaning of facial expressions. *Current Directions in Psychological Science*, *27*(4), 211–219. https://doi.org/10.1177/0963721417746794

Godoy, R., Reyes-García, V., Huanca, T., Tanner, S., Leonard, W. R., McDade, T., & Vadez, V. (2005). Do smiles have a face value? Panel evidence from Amazonian Indians. *Journal of Economic Psychology*, *26*(4), 469–490. https://doi.org/10.1016/j.joep.2004.10.004

Götz, F. M., Stieger, S., & Reips, U.-D. (2017). Users of the main smartphone operating systems (iOS, Android) differ only little in personality. *PLOS ONE*, *12*(5), Article e0176921. https://doi.org/10.1371/journal.pone.0176921

Grammatikopoulou, A., Grammalidis, N., Bostantjopoulou, S., & Katsarou, Z. (2019). Detecting hypomimia symptoms by selfie photo analysis: For early Parkinson disease detec-

tion. *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 517–522. https://doi.org/10.1145/3316782.3322756

Green, K. (2013). *On Fire* [Comic]. http://gunshowcomic.com/648

Kachur, A., Osin, E., Davydov, D., Shutilov, K., & Novokshonov, A. (2020). Assessing the Big Five personality traits using real-life static facial images. *Scientific Reports*, *10*(1), Article 8487. https://doi.org/10.1038/s41598-020-65358-6

Lee, E.-H. (2012). Review of the Psychometric Evidence of the Perceived Stress Scale. *Asian Nursing Research*, *6*(4), 121–127. https://doi.org/10.1016/j.anr.2012.08.004

Lind, M. N., Byrne, M. L., Wicks, G., Smidt, A. M., & Allen, N. B. (2018). The Effortless Assessment of Risk States (EARS) tool: An interpersonal approach to mobile sensing. *JMIR Mental Health*, *5*(3), Article 3. https://doi.org/10.2196/10334

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), Article 3. https://doi.org/10.1016/0005-7967(94)00075-U

Mayo, L. M., & Heilig, M. (2019). In the face of stress: Interpreting individual differences in stress-induced facial expressions. *Neurobiology of Stress*, *10*, Article 100166. https://doi.org/10.1016/j.ynstr.2019.100166

Qiu, L., Lu, J., Yang, S., Qu, W., & Zhu, T. (2015). What does your selfie say about you? *Computers in Human Behavior*, *52*, 443–449. https://doi.org/10.1016/j.chb.2015.06.032

Rychlowska, M., Jack, R. E., Garrod, O. G. B., Schyns, P. G., Martin, J. D., & Niedenthal, P. M. (2017). Functional smiles: Tools for love, sympathy, and war. *Psychological Science*, *28*(9), 1259-1270. https://doi.org/10.1177/0956797617706082

Siegal, J. (2017, July 20). Android continues to increase its sizable lead over iOS in the US. *BGR*. https://bgr.com/tech/android-vs-ios-market-share-2017-q2/

Slavich, G. M., & Shields, G. S. (2018). Assessing lifetime stress exposure using the Stress and Adversity Inventory for Adults (Adult STRAIN): An overview and initial validation. *Psychosomatic Medicine*, *80*(1), 17–27. https://doi.org/10.1097/PSY.0000000000000534

Zunhammer, M., Eberle, H., Eichhammer, P., & Busch, V. (2013). Somatic symptoms evoked by exam stress in university students: The role of alexithymia, aeuroticism, anxiety and depression. *PLOS ONE*, *8*(12), Article e84911. https://doi.org/10.1371/journal.pone.0084911

# A Smiling Paradox: Exploring the Constructed Nature of Emotions. A Reflection on the Relationship Between Smiling in Selfies and Distress.

Anne Margit Reitsema[1,2,3], Sanne Nijhof[2,3], Odilia Laceulle[1,2]

**Keywords** *facial expressions, emotion, selfie, emotion recognition*

[1] Department of Developmental Psychology, Utrecht University

[2] Research Theme Dynamics of Youth, Thriving & Healthy Youth, Utrecht University

[3] Department of Pediatrics, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht University

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
July 15, 2024
**Accepted**
October 15, 2024
**Published**
December 11, 2024

**Correspondence**
Utrecht University
a.m.reitsema@uu.nl



Much has been written about facial expressions and emotions, with numerous references pointing back to the world-famous studies conducted by Paul Ekman on universal expressions of emotion (see Figure 1). The work of Ekman and colleagues famously showed that even people from remote tribes in Papua New Guinea could recognize emotions from pictures of facial expressions (Ekman & Friesen, 1969). This fostered the general belief that certain basic emotions – six to be precise – are part of our nature and shared universally. The study "Smile you're on camera: Investigating the relationship between selfie smiles and distress" (Lind et al., 2024) is one of many recent studies that challenge this fundamental emotion perspective. It urges us to consider a completely opposite view: Emotions are not universal but constructed through our cultural and personal experiences. Moreover, the study raises important questions on how much value society places on being – or appearing – happy, and the promises of AI powered emotion detection.

**Figure 1** Photographs used in cross-cultural research by Ekman and Friesen (1969).

## The Study

In the current study (Lind et al., 2024), participants were asked to record daily two-minute videos in which they described one positive and one negative event from their day. Importantly, these videos were recorded in selfie

mode, which means participants viewed themselves during the recording. The researchers subsequently used an automated facial behavior analysis toolkit to analyze the facial expressions in the videos. They focused on one specific so-called facial action unit that is commonly referred to the "lip corner puller", also shown in the first photograph in Figure 1. While the researchers expected that more intense smiling would correlate with lower distress, their results showed the opposite pattern: Smiling intensity was associated with *higher* levels of anxiety, depression, and stress. Although this may seem counterintuitive, these findings align with a growing body of research that shows that emotions are not universally and consistently experienced and expressed (Barrett et al., 2019). Instead, emotions might be *constructed* based on a variety of factors, including past experiences and cultural norms.

## Emotions in Context

The study's results, as the authors themselves also note, challenge Ekman's idea of discrete and universal "basic emotions". We think it is valuable to reflect on this a bit deeper, since this idea still features prominently in many psychology textbooks and the popular press (e.g., American Psychological Association, 2018). A number of studies have now failed to replicate Ekman's findings (Barrett et al., 2019). The original Ekman study relied on forced-choice methods, which might have biased the results by priming participants with predetermined emotion categories. In studies that used a greater diversity of research methods, such as free sorting, people from remote tribes did not label the facial expressions as Ekman and his colleagues proposed, performing no better than chance (Gendron et al., 2018). These newer studies indicate that contextual factors play a much larger role in emotion perception than previously thought.

According to the theory of constructed emotion (Barrett, 2017), emotions are not pre-set reactions that we are born with but are created by our brains on the spot by using a combination of bodily sensations, contextual information, and past experiences. To maximize efficiency and minimize energy use, our brains create emotions by using both past experiences and the current situation to predict to new situations. Instead of processing each experience "from scratch", the brain efficiently reuses existing knowledge. This means that our emotions are very context-dependent: The same situation can evoke different emotions depending on an individual's prior experiences and the current environment.

A range of studies support this constructed view of emotions by showing that things like facial expressions, autonomic responses, and brain activity do not form distinct, consistent clusters that can clearly differentiate one type of emotion from another (e.g., Lindquist et al., 2012; Siegel et al., 2018). Instead, variability is the norm. The same emotion can lead to different facial expressions depending on cultural expectations and personal history. For instance, when you are angry, your facial expression is not always the same and does not always look exactly as the expression in the fourth photograph in Figure 1. Sometimes you might frown when you are angry, other times you may cry, or you might show no obvious sign of anger at all if it is not appropriate for the situation. The expressions in Figure 1 are therefore better thought of as stereotypes that fail to capture the rich variety of emotion expressions that likely emerge in daily life.

Applying the theory of constructed emotions to the current study provides valuable insights into the unexpected association between smiling intensity and higher distress. In the unique context of recording selfie videos, participants may construct emotions and expressions in ways that differ from everyday life. Participants that experience higher levels of anxiety or stress might construct their emotional response to include more intense smiling as a way to cope with or regulate their internal discomfort. Moreover, smiling intensely in this context could itself potentially contribute to feelings of distress, as participants become more self-aware and perhaps self-critical. This highlights the theory's core idea that emotional experiences and expressions are deeply shaped by context—what might appear to be a smile of happiness may, in fact, be a response to stress or anxiety.

## Social Norms and Emotion

Social norms play a crucial role in shaping how emotions are expressed and interpreted. This

is particularly relevant in the context of the current study, where participants were continually confronted with their own expressions while recording selfie videos. Lind and colleagues discussed this in the context of Fridlund's behavioral ecology view (1994), which proposes that facial expressions are better understood as tools for communication rather than direct reflections of a person's inner emotions. These communicative tools are heavily influenced by cultural norms, which dictate which emotions should be expressed or suppressed. Given that participants were more self-aware while recording themselves, they might have been influenced by these cultural norms, leading them to display emotions that do not necessarily reflect their true feelings.

In many cultures, especially in Western societies, there is an emphasis on happiness and the maximization of positive emotions and minimization of negative ones (Bastian et al., 2015). There is often a societal expectation to always look happy, which can lead to frequent smiling and acting cheerful even when people do not feel that way, which ironically is linked to poorer well-being (Dejonckheere et al., 2022). This trend is particularly strong on social media where people predominantly share positive moments and emotions and downplay or omit negative ones. This can make the pressure to fit in with societal expectations even stronger. Moreover, the constant exposure to such idealized displays of happiness can lead to unrealistic social comparisons and negatively impact mood (Bennett et al., 2017).

In the context of the original study, the societal expectation to display positive emotions, such as smiling, may be amplified by the artificial setting. The act of recording a selfie video may trigger awareness of social expectations, particularly the pressure to appear positive in self-presentations. This could lead participants to smile more intensely to conform to these social norms. Given our growing understanding of emotions as social and contextualized phenomena, future research should prioritize methods that capture this complexity. This could include, for example, more ecological momentary assessment research to study emotions in real-time natural settings, conducting cross-cultural studies to examine how social norms influence emotion construction and expression, and developing multimodal assessment techniques that combine facial expressions, physiological measures, and self-reports.

## Practical and Ethical Considerations

Facial expression analysis, like the kind used in this study, is rapidly gaining popularity and is increasingly applied in many fields, including psychology. It is one of the primary methods used in emotion-detection technology, alongside voice analysis, gait analysis, and eye tracking. Many companies, including Affectiva (an MIT spin-off) and Hume AI (which recently received $50 million in new funding, Business Wire, 2024), as well as tech giants like Apple, Google, and Facebook, are actively developing these technologies. These innovations are believed to provide companies with a competitive advantage and are already being implemented in diverse areas. For instance, they are used during job interviews to evaluate candidates' emotional responses (Unilever; Hymas, 2019), to predict the popularity of movies (Disney; Deng et al., 2017), and even for personalized menu recommendations in fast food (KFC; Hawkins, 2017).

However, the keyword in all of this is *accuracy*. The basis for emotion-recognition technology is just not there. After reviewing over 1,000 papers on facial expression of emotion, Barrett and colleagues (2019, p. 46) concluded: "It is not possible to confidently infer happiness from a smile, anger from a scowl, or sadness from a frown, as much of current technology tries to do when applying what are mistakenly believed to be the scientific facts." This is similar to the flawed logic behind polygraphs or "lie detectors": The assumption that we can directly infer people's mental states from their physical expressions. While these technologies can be very accurate at detecting facial expressions, they ultimately cannot tell us what these expressions mean or what the person is actually feeling.

Will it then ever be possible to have accurate AI emotion detection? Maybe, but there are significant challenges to overcome. Right now, most AI technologies rely on simplified facial expressions to detect emotions, which – as we saw – does not capture the full picture (Barrett et al., 2019; Cabitza et al., 2022; Cross et al., 2023). For AI to get better at "reading" emo-

tions, it needs to move past these stereotypes and recognize the variety of possible expressions and the context in which they occur. This means that other factors also need to be considered, like environment and time of day. For any AI to detect emotions accurately, it must understand the context in which expressions occur, as emotions are context-bound (Barrett, 2022). Detailed data from people in various real-life situations can allow for a more complete understanding of emotional expression. This involves using multiple types of data, like voice analysis and body posture, to get a fuller picture. By doing this, AI could better understand that a smile does not always mean someone is happy – they might be smiling but in reality, feel stressed or upset.

## Conclusion

The study's finding that intense smiling can signal distress highlights the complexity of emotions. It illustrates that expressions are influenced by context, shaped by social norms, and deeply intertwined with personal experiences. These findings challenge the traditional view that facial expressions are straightforward reflections of internal states. They also raise crucial questions for AI emotion-detection systems, which often simplify the interpretation of expressions. Moving forward, research should prioritize exploring emotions in real-world settings, using more refined methods that can capture their variability and complexity. By advancing our understanding of how context shapes emotional expression, both research and AI systems can move toward more accurate and holistic interpretations of emotional signals.

## References

American Psychological Association. (2018). *Primary emotion*. In *APA Dictionary of Psychology*. Retrieved July 15, 2024, from https://dictionary.apa.org/primary-emotion

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience, 12*, 1–23. https://doi.org/10.1093/scan/nsw154

Barrett, L. F. (2022). Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *American Psychologist*, *77*(8), 894-920. https://doi.org/10.1037/amp0001054

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1-68. https://doi.org/10.1177/1529100619832930

Bastian, B., Koval, P., Erbas, Y., Houben, M., Pe, M., & Kuppens, P. (2015). Sad and alone: Social expectancies for experiencing negative emotions are linked to feelings of loneliness. *Emotion, 6*(5), 496–503. http://doi.org/10.1177/1948550614568682

Bennett, B. L., Whisenhunt, B. L., Hudson, D. L., Wagner, A. F., Latner, J. D., Stefano, E. C., & Beauchamp, M. T. (2020). Examining the impact of social media on mood and body dissatisfaction using ecological momentary assessment. *Journal of American College Health*, *68*(5), 502-508. https://doi.org/10.1080/07448481.2019.1583236

Business Wire. (2024, March 27). *Hume AI announces $50 million fundraise and empathic voice interface*. Business Wire. https:/ / www .businesswire.com/ news/ home/ 20240326359639/ en/ Hume-AI-Announces-50-Million-Fundraise-and-Empathic-Voice-Interface

Cabitza, F., Campagner, A., & Mattioli, M. (2022). The unbearable (technical) unreliability of automated facial emotion recognition. *Big Data & Society*, *9*(2). https://doi.org/10.1177/20539517221129549

Cross, M. P., Acevedo, A. M., & Hunter, J. F. (2023). A critique of automated approaches to code facial expressions: What do researchers need to know? *Affective Science*, *4*(3), 500-505. https://doi.org/10.1007/s42761-023-00195-0

Dejonckheere, E., Rhee, J. J., Baguma, P. K., Barry, O., Becker, M., Bilewicz, M., Castelain, T., Costantini, G., Dimdins, G., Espinosa, A., Finchilescu, G., Friese, M., Gastardo-Conaco, M. C., Gómez, A., González, R., Goto, N., Halama, P., Hurtado-Parrado, C., Jiga-Boy, G. M., ... Bastian, B. (2022). Perceiving societal pressure to be happy is linked to poor well-being, especially in happy nations. *Scientific Reports*, *12*(1), Article 1514. https:/ / doi.org/ 10.1038/ s41598-021-04262-z

Deng, Z., Navarathna, R., Carr, P., Mandt, S., Yue, Y., Matthews, I., & Mori, G. (2017). Fac-

torized variational autoencoders for modeling audience reactions to movies. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2577-2586). https://doi.org/10.1109/CVPR.2017.637

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*(3875), 86-88. https://doi.org/10.1126/science.164.3875.86

Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. Academic Press.

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality reconsidered: Diversity in making meaning of facial expressions. *Current Directions in Psychological Science*, *27*(4), 211-219. https://doi.org/10.1177/0963721417746794

Hawkins, A. (2017, January 11). *KFC China is using facial recognition tech to serve customers – but are they buying it?* The Guardian. https://www.theguardian.com/technology/2017/jan/11/china-beijing-first-smart-restaurant-kfc-facial-recognition

Hymas, C. (2019, September 27). *AI facial recognition used for first time in job interviews in UK to find best applicants*. The Telegraph. https://www.telegraph.co.uk/news/2019/09/27/ai-facial-recognition-used-first-time-job-interviews-uk-find/

Lind, M. N., Byrne, M. L., Devine, S., & Allen, N. B. (2024). Smile, you're on camera: Investigating the relationship between selfie smiles and distress. *Journal of Trial & Error*. https://doi.org/10.36850/8716-5abe

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences*, *35*(3), 121-143. https://doi.org/10.1017/S0140525X11000446

Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., Quigley, K. S., & Barrett, L. F. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychological Bulletin*, *144*(4), 343-393. https://doi.org/10.1037/bul0000128

**Correction notice**
Incorrect Special Issue Labeling (Article erroneously excluded): This article was previously not labeled as part of a special issue due to an error. This has now been corrected.

# Prenatal Sildenafil and Fetal-placental Programming in Human Pregnancies Complicated by Fetal Growth Restriction: A Retrospective Gene Expression Analysis

Fieke Terstappen[1], Torsten Plösch[2], Jorg J.A. Calis[3,4], Wessel Ganzevoort[5], Anouk Pels[5], Nina D. Paauw[1], Sanne J. Gordijn[2], Bas B. van Rijn[6], Michal Mokry[3], A. Titia Lely[1]

[1]University Medical Center Utrecht, Wilhelmina Children's Hospital, Department of Obstetrics, Utrecht, The Netherlands

[2]University Medical Center Groningen, Department of Obstetrics and Gynaecology, University of Groningen, Groningen, The Netherlands.

[3]University Medical Center Utrecht, Department of Cardiology, Utrecht, The Netherlands.

[4]University Medical Centre Utrecht, Center for Translational Immunology, Utrecht, The Netherlands

[5]Amsterdam University Medical Centers, Department of Obstetrics, University of Amsterdam, Amsterdam, The Netherlands

[6]Erasmus MC University Medical Center Rotterdam, Department of Obstetrics and Fetal Medicine, Rotterdam, The Netherlands

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
March 30, 2023
**Accepted**
August 23, 2023
**Published**
September 26, 2023

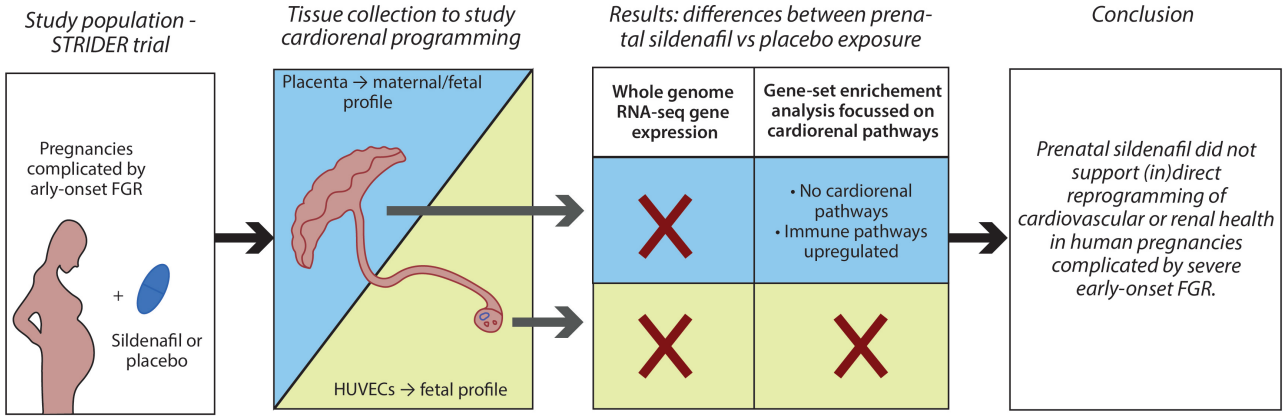
**Correspondence**
University Medical Center Utrecht
F.Terstappen@umcutrecht.nl



**Objective:** Fetal growth restricted (FGR) offspring are more susceptible to develop cardiovascular and renal disease. The potential therapeutic value of sildenafil to improve fetal growth has recently been evaluated in several randomized clinical trials. Here we investigate whether administration of sildenafil during pregnancies complicated by FGR influences fetal-placental programming profiles, especially related to cardiorenal development and disease.
**Methods:** We collected human umbilical vein endothelial cells (HUVECs) and placental tissue within the Dutch STRIDER trial, in which sildenafil versus placebo treatment were randomly assigned to pregnancies complicated by severe early-onset FGR. Differential expression of genes of these samples were studied by whole genome RNA-sequencing. In addition, we performed gene set enrichment analysis focused on cardiovascular and renal gene sets to examine differentially expressed gene sets related to cardiorenal development and health.
**Results:** Our study showed similar gene expression profiles between treatment groups in HUVECs (n=12 sildenafil; n=8 placebo) and placentas (n=13 per group). Prenatal sildenafil exposure did not change cardiovascular or renal programming in pregnancies complicated by FGR. In placental tissue, prenatal sildenafil altered a few gene sets involved with the nitric oxide pathway potentially reflecting the mechanism of action of sildenafil. Prenatal sildenafil also upregulated gene sets related to immune pathways in placental tissue.
**Conclusions:** Overall, our study showed that sildenafil has the potential to alter placental (but not fetal) expression of gene sets related to immune pathways and did not support (in)direct reprogramming of cardiovascular or renal health in human pregnancies complicated by FGR.

**Keywords** *fetal growth restriction, sildenafil, developmental programming, RNA-sequencing, gene set enrichment analysis, human umbilical cord vein endothelial cells, placenta*

Fetal growth restriction (FGR) describes the inability of the fetus to reach its intrinsic growth potential. Early-onset FGR is most commonly caused by maternal vascular malperfusion resulting in placental insufficiency; a reduced uteroplacental blood flow impairs oxygen and nutrient supply towards the fetus s (Burton & Jauniaux, 2018). Multiple lines of evidence suggest that exposure to this unfa-




Graphical abstract

## Take-home message

Fetal growth restriction (FGR) predisposes to cardiorenal diseases later in life. Prenatal administration of sildenafil showed beneficial effects on cardiovascular health in animal FGR offspring. However, this study did not support the potency of sildenafil to reprogram the cardiorenal health in human pregnancies complicated by FGR.

vorable intrauterine environment results in fetal adaptations, impaired maturation of organ development, and triggers developmental programming of cardiorenal diseases later in life (Barker, 2006; Sundrani et al., 2017; Chen & Zhang, 2012; Henriksen & Clausen, 2002). FGR offspring, and their future generations, are more susceptible to develop cardiovascular and renal disease, including hypertension, ischemic heart disease, chronic kidney disease and end-stage renal disease in adulthood (Malhotra et al., 2019; White et al., 2009; Gjerde et al., 2020; Kooiman et al., 2020; Demicheva & Crispi, 2014; Dötsch et al., 2016; Sehgal et al., 2020; Nüsken et al., 2020). Adverse developmental programming might underlie this increased risk (Sehgal et al., 2020; Nüsken et al., 2020). A recent study showed altered gene expression and gene sets related to cardiovascular health and renal development in human umbilical vein endothelial cells (HUVECs) between placental insufficiency-induced FGR and control (Terstappen et al., 2020).

Mitigation of the detrimental developmental programming of cardiorenal disease following FGR in pregnancy is desirable e (Paauw et al., 2016). Several animal studies provided proof-of-concept of *in utero* interventions to diminish adverse developmental programming, also known as reprogramming. For instance administration of nitric oxide (NO) stimulating agents during pregnancy improved cardiovascular outcomes alongside altered epigenetics and gene expression in fetuses or offspring of FGR models in rat, guinea pig or chicken (Herrera et al., 2017; Itani et al., 2017; Man et al., 2020; Z. Wu et al., 2015).

Sildenafil is one of these potential candidates to improve fetal growth and reprogram cardiorenal disease. Prenatal administration of this phosphodiesterase-5 showed beneficial effects on cardiovascular function in FGR models in rats, mice, and chickens (Itani et al., 2017; Mills et al., 2018; Terstappen et al., 2019). This might be a result of indirect influence on developmental programming, via improved fetal growth by uteroplacental blood flow by NO induced relaxation of the placental vascular bed on both the maternal and fetal side (Krause et al., 2011; Wareing et al., 2006).. In addition, the long-term benefit might also result from direct protection of developmental programming, since improvements in long-term cardiovascular function were also observed in a chick embryo model (without the presence of a pla-

centa) of FGR after administration of sildenafil (Itani et al., 2017). While several animal and small-scaled human studies indeed showed improved fetal growth in the sildenafil-treated group (Paauw et al., 2017; Von Dadelszen et al., 2011), a series of randomized clinical trials (Sildenafil TheRapy In Dismal Prognosis Early-onset Fetal Growth Restriction [STRIDER]) designed within an international network did not show increased birth weight in pregnancies complicated with severe early-onset FGR following sildenafil compared to placebo (Pels et al., 2020; Sharp et al., 2018; Groom et al., 2019). These human data make indirect influence on developmental programming less likely; nevertheless, prenatal sildenafil could still directly influence developmental programming.

We collected HUVECs and placental tissue as a sub-study within the Dutch STRIDER trial to study whether sildenafil influences programming by regulation of fetal and placental gene expression. Hereby we aimed to address potential mechanisms underlying the effects of sildenafil on cardiovascular and renal programming in severe early-onset FGR offspring. Our approach involved whole genome RNA-sequencing to map differential expression per gene and gene set enrichment analysis focussed on cardiovascular and renal development, function and health.

## Methods

### Study population

For this substudy, women with a pregnancy complicated by severe early-onset FGR most likely based on placental insufficiency and participating in the Dutch STRIDER were recruited from the University Medical Center Utrecht (UMCU) and the Academic Medical Center (AMC) in Amsterdam from 09-07-2016 to 30-10-2017. In the Dutch STRIDER study, women were randomized to either prenatal administration of placebo or sildenafil at a dose of 25mg three times per day (Pels et al., 2020; Pels et al., 2017). The study population was selected based on low biometric parameters for gestational age and signs of placental insufficiency; inclusion and exclusion criteria have been described in detail previously (Pels et al., 2020; Pels et al., 2017). For this sub-study, we excluded cases in whom the offspring was diagnosed with a congenital disorder after birth.

The Dutch STRIDER trial (Clincial trial.gov identifier NCT02277132) was approved by the medical ethical committee of the AMC on 02-07-2014; protocol number 2014-131. The UMCU approved the study on 14-09-2015; protocol number 15-510/G-C. Prior to delivery, the STRIDER participants in this sub-study gave additional written informed consent for placental research (amendment approved on 29-01-2016, updated on 05-09-2017). The Dutch STRIDER study was halted mid-term because the interim-analysis showed futility in combination with potentially increased risk of mortality and persistent pulmonary hypertension (PPHN) in neonates.

### Clinical data

Clinical data has been derived from the Dutch STRIDER database and patient records. The maternal medication was noted, including the start and duration of administration of the allocated drug. Unblinding of treatment allocation was done after all methods below were executed. Exact percentiles for weight and head circumference at birth were determined with Intergrowth-21$^{st}$ (Anderson et al., 2016). Neonatal survival during hospital admission and cases of PPHN were registered.

### HUVECs isolation and RNA isolation

Directly after placental delivery, the umbilical cord was stored in phosphate-buffered saline (PBS) solution (pH 7.2) at 4°C. HUVECs isolation occurred as previously described, preferably within 12 hours, but always within 24 hours after placental delivery (Hartman et al., 2020). We collected umbilical cords in both UMCU and AMC. Cannulation of the umbilical vein at one end allowed access for further processing. After washing with sterile PBS (pH 7.4; Gibco by Life Technologies, Grand Island, NY) the umbilical cord was clamped at both sides to incubate with accutase (0.02 µg/ml DNase; Innovative cell technologies Inc, San Diego, CA) for 5 minutes in 37°C sterile PBS to detach the endothelial cells. The detached HUVECs in accutase were flushed out of the umbilical vein with endothelial cell growth medium-2 (97% EGM-2; basal medium and SingleQuots supple-

ment [1.9% FBS, 0.04% hydrocortisone, 0.4% hFGF-B2, 0.1% vascular endothelial growth factor, 0.1% R3-IGF-1, 0.1% ascorbic acid, 0.1% hEGF, 0.1% GA-1000, 0.1% heparin], Lonza Bioscience, Walkersville, MD) and centrifuged for 5 minutes at 330g at room temperature. The pellet was resuspended in 600 µl RA1 lysis buffer (Macherey-Nagel, Düren, Germany) and 6 µl 1M DTT and stored at -80°C until RNA isolation.

RNA was isolated using NucleoSpin RNA® (Macherey-Nagel), with RNA elution in 40 µl nuclease-free water. RNA concentration was quantified using Qubit RNA HS assay and Qubit fluorometer (ThermoFisher).

### RNA isolation of placental tissue

Biopsies (4 by 4 mm) were taken from the middle of five cotyledons per placenta directly after birth and only in AMC. RNAlater stabilization solution (ThermoFisher Scientific) was used to freeze the (unrinsed) placental biopsies in liquid nitrogen and stored at -80°C until RNA isolation.

Frozen placenta samples were homogenized with a Tissuelyzer LT (Qiagen, Venlo, the Netherlands) in lysis buffer. Total RNA was isolated with the allprepRNA mini kit (Qiagen), following the manufacturer's protocol. RNA quality and quantity were characterized by a Nanodrop 2000c (Thermo Scientific, Pittsburgh, PA, USA). RNA was stored at -80°C until further analysis.

### RNA-sequencing of HUVECs and placental tissue

Samples with a measurable amount (minimal concentration 41.4 ng/µl in placental samples and 2.1 ng/µl in HUVEC samples) of RNA were selected for RNA sequencing. Polyadenylated mRNA was isolated using Poly(A) beads (NEXTflex). Sequencing libraries were prepared by using the Rapid Directional RNA-seq kit (NEXTflex). The library was sequenced at the Utrecht Sequencing Facility (USEQ) on a Nextseq500 platform (Illumina) using a single-end 75-base pair high-output run. Reads were aligned to the human reference genome (GRCh37) using STAR version 2.4.2a. Read groups were added to the BAM files with Picard's AddOrReplaceReadGroups (v1.98). The BAM files were sorted with Sambamba v0.4.5, and transcript abundances were quantified with HTSeq-count version 0.6.1p117 using the union mode.

### Gene set analysis

Gene set enrichment testing was performed on the hallmark (H), canonical pathway (C2-CP) and GO term (C5) gene set collections from the Molecular Signatures Database (version 7.1) (D. Wu & Smyth, 2012; Liberzon et al., 2016). Only gene sets with relation to renal or cardiovascular development or pathologies, or Nitric Oxide (NO) signaling were selected from the GO term gene sets (**Table S1**). Gene sets with less than five genes in the set of selected genes (based on expression, see above) were excluded from the analysis, eventually resulting in 2,167 included gene sets

### Statistical analysis

*Clinical data*

IBM SPSS Statistics 25 for Windows (version 25, UBM Corp, Armonk, NY) was used for statistical analysis. Parametric data tested with independent t-test are presented as mean ± SD, non-parametric data tested with Mann-Whitney are presented as median (minimum-maximum), and nominal data tested with Fisher exact are presented as n (%). A two-sided p-value of below 0.05 was considered statistically significant.

*Differential expression of genes*

Read counts per gene, per sample, were analyzed for global expression differences using R (version 3.5.3). Genes were selected with an expression of one count per million reads (CPM) in at least 8 samples (n=13,760 genes selected). Read counts were Trimmed Mean of M-values (TMM)-normalized using the calcNormFactors function from the edgeR package (version 3.24.3) (Robinson et al., 2010). TMM-normalized counts were used to assess global transcriptional profile differences of all samples by Principal Comonent Analysis (PCA) (ten components). Ten Principal components (PC) were analyzed in the PCA analysis, values from each PC were checked for correlation to sample characteristics by the Mann-Whitney U test implemented in the scipy package (version 0.19.0) in python (version 2.7.10). Low-quality

samples were identified and removed when passing any of these conditions: 1) number of reads were less than 1,000,000, 2) number of non-zero genes were less than 10,000, or 3) a combination of number of non-zero genes between 10,000-12,000 and being a visible outlier on one of the PCA components. HUVECs and placental tissue were analyzed separately.

Differential gene expression analysis was performed with the edgeR package (version 3.24.3) in R (version 3.5.3, R Core team, Auckland, New Zealand). Gene expression was modeled using the glmQLFit function in EdgeR (Robinson et al., 2010), to a model that included treatment group variables, as well as factors to capture mode of delivery (caesarean section vs spontaneous delivery), sex (male vs female), and gestational stage (preterm vs term) related gene expression variation. Differential gene expression was determined between treatment groups (sildenafil vs placebo) and the differential expression statistics were obtained using the glmQLFTest functionality in edgeR. False Discovery Rates (FDR) were determined using the Benjamini-Hochberg method to adjust for multiple testing and were considered significant when below 0.1 (in combination with unadjusted p-value <0.05; (Benjamini & Hochberg, 1995).

*Differential expression of gene sets*

Gene set enrichment testing was performed with Correlation Adjusted MEan Rank (CAMERA), using the same linear model and contrasts as in the differential gene expression analysis (see above), and FDR were determined using the Benjamini-Hochberg method to adjust for multiple testing, which were considered significantly different when below 0.1 (Benjamini & Hochberg, 1995). When a module showed ≥ 50% overlap with higher ranking gene sets we only selected the more significant gene set. Heatmap for the gene sets related to the cardiovascular, renal and nitric oxide pathway were created.

## Results

### Sample inclusion

We collected umbilical cords from 14 sildenafil-treated and 10 placebo-treated births. Two sildenafil-treated and one placebo-treated sample did not yield enough RNA to perform RNA-sequencing. Therefore 12 sildenafil and 9 placebo HUVECs samples were used for RNA-sequencing. From the HUVECs samples, we excluded 1 low-quality run from the placebo group that was also an outlier in the RNA-seq data. Therefore, we proceeded with analyzing 12 sildenafil-treated versus 8 placebo-treated HUVEC samples.

We collected 16 sildenafil-treated and 18 placebo-treated placental biopsies. The RNA concentration of one placebo sample was too low to perform RNA-sequencing and thereby excluded. We also excluded one placental tissue sample in the placebo group based on postnatal detection of congenital disorders (Silver Russell syndrome). This resulted in 16 sildenafil versus 16 placebo-treated placental samples for RNA-sequencing. In these results, we identified three low-quality samples outliers in both the sildenafil and placebo group. Thus, we proceeded with 13 sildenafil-treated and 13 placebo-treated placenta samples for analysis.

### Study characteristics

Patient characteristics are presented in **Table 1**. Birth weight did not differ between groups. Neonatal survival during hospital admission was approximately 20% higher in the sildenafil group compared to the placebo group (not significantly different). Only sildenafil-exposed neonates suffered from PPHN during admission (n=3 in HUVECs samples and n=2 in the placental tissue samples).

### Differential expression of genes

Principal component analysis (PCA) plots did not reveal clear clustering of sildenafil versus placebo in HUVECs (**Figure 1A**) or placental tissue (**Figure 1B**). The heatmaps supported that the gene expression between samples were similar (**Figure S1**). Multidimensional scaling (MDS) plots of HUVECs material showed clustering in prematurity as potential modifier, but not in the treatment group, delivery route, or sex (**Figure S2**). MDS plots of placental material showed no clustering in the treatment group, sex, prematurity or delivery route as potential modifiers (**Figure S3**). All of the study characteristics were tested for association for all of the first 10 PCs. Gestational stage was as-

sociated with PC2 and PC5 in HUVECs and PC8 in placenta, delivery route was associated with PC2 and PC5 in HUVECs and PC1 and PC10 in placenta, and sex was associated with PC3 in HUVECs and PC8 in placenta (**Table S2**). Therefore, differences in expression due to mode of delivery, gestational stage, and sex were accounted for in the modeling of gene expression.

The analysis of differential expression of genes, including genes involved in the NO pathway or related to cardiovascular or renal development or function, did not show any significant differences between the treatment groups, neither in HUVECs samples (**Table S3**) nor in placenta samples (**Table S4**).

### Differential expression of gene sets

Gene set enrichment analysis did not show any differences between treatment groups in HUVECs samples (**Table S5**). However, in placental samples 90 gene sets were upregulated in sildenafil-treated compared to placebo-treated (**Table S6**). The selection of only the highest-ranking gene set module (overlapping modules excluded) resulted in 64 upregulated gene sets and 5 downregulated gene sets. These gene sets mostly involved immune pathways, and three gene sets were related to the NO pathway and one to cardiovascular disease (**Table 2**). Heatmaps were made for the top ten gene sets related to immune pathways (**Figure S4**), and the four gene sets related to the NO pathway and cardiovascular disease (**Figure S5**). This was done to study the extent of up- and downregulation for the distinct genes in these gene sets in each sample and were ordered per duration of treatment. From this analysis, most genes were up- and downregulated in accordance with the differential gene set analysis results for immunity, but this did not apply for gene sets related to the NO pathway and cardiovascular disease. None of the heatmaps showed ascending or descending expression levels correlating to duration of treatment.

The adverse *in utero* environment resulting in fetal growth restriction (FGR) predisposes the offspring to develop cardio-renal disease beyond the fetal developmental phase by altered epigenetic programming. Interest grew in prenatal administration of sildenafil after several animal studies showed improved fetal growth and long-term cardiovascular function.

## Discussion

This sub-study within the Dutch STRIDER trial evaluated whether prenatal sildenafil administration during pregnancies complicated by severe early-onset FGR influenced gene modules, with specific focus on cardiovascular and renal programming. The RNA expression in collected HUVECs and placental tissue did not differ between the sildenafil or placebo group. Gene set enrichment analysis also showed no differences in gene sets related to cardiovascular and renal health in HUVECs, but three gene sets involved in the NO-pathway and one in cardiovascular health were possibly different in placenta samples. Additionally, we observed an upregulation of several gene sets related to immune pathways in the sildenafil-exposed placental samples.

### Lack of difference in gene expression and gene sets related to cardiorenal health

The results of the three STRIDER trials in UK, New-Zealand and the Netherlands suggested that prenatal sildenafil does not have an indirect programming effect via placental function improvements since they observed no beneficial effects on pregnancy outcomes, such as birth weight, prolongation of pregnancy, or perinatal morbidity or mortality (Groom et al., 2019; Sharp et al., 2018; Pels et al., 2020). Complementary to this, the results from this current sub-study showed no direct cardiovascular or renal programming following prenatal sildenafil exposure. The low sample size due to the mid-term halt of the study and only collecting samples in part of the participating centers due to logistics might both have limited detection of beneficial effects. Alternatively, our lack of clear results regarding cardiovascular and renal programming might be a result of interspecies differences, since several animal studies did report a reprogramming potential of prenatal administration of NO-stimulating agents (sildenafil, pentaerythritol tetranitrate, N-acetely cysteine) in animal models for placental insufficiency (Herrera et al., 2017; Itani et al., 2017; Z. Wu et al., 2015). However, a recent study showed that prenatal sildenafil re-

programmed salt-sensitive hypertension in rat FGR offspring, but had not affected renal function nor did they find differences in targeted RNA-seq data in renal tissue (Turbeville et al., 2020). These conflicting results might plead for the use of tissue collected from complicated pregnancies (such as FGR) in humans to study developmental programming on a molecular level rather than the use of animal tissue.

### Upregulated gene sets involved with NO pathway

We observed upregulation of three gene sets related to the NO pathway (response to *v*ascular endothelial growth factor stimuli, leukocyte adhesion to vascular endothelial cells, and negative regulation of the NO metabolic process) in placental tissues in the sildenafil group compared with the placebo group. These gene sets might reflect the mechanism of action of sildenafil. Preclinical studies report conflicting results, with some showing altered expression of metabolites in the NO-pathway after sildenafil exposure in fetal cardiac or lung tissue (Itani et al., 2017; Shue et al., 2014), while others did not find these differences in expression in spite of beneficial functional effects (George et al., 2013). However, while these gene sets were significantly different in our study, the heatmaps of these gene sets showed that the differences of the individual genes were minimal and independent of duration of sildenafil intake. Therefore, these inconclusive results did not lead to a clear insight regarding the mechanism of action of sildenafil.

The Dutch STRIDER trial showed a potentially increased risk of PPHN in neonates (Pels et al., 2020). The gene set of increased leukocyte adhesion to vascular endothelial cells combined with our gene set's result on immune pathways might suggest that this pathway is involved in the increased PPHN risk (Rafikov et al., 2019; Kuebler et al., 2018; El Chami & Hassoun, 2012; Kobayashi et al., 2004). However, this was not observed in HUVECs representing the fetal profile. To gain better insight into underlying mechanisms involved with the potentially increased risk of PPHN in the Dutch STRIDER trial requires follow-up studies that were beyond the scope of this study.

### Upregulated gene sets involved in the immune pathway

Interestingly, sildenafil administration resulted in the upregulation of several gene sets involved in immune or inflammation pathways and, additionally, longer sildenafil intake correlated with higher expression of genes related to these pathways. Pregnancies complicated by placental insufficiency syndromes show an increased placental release of pro-inflammatory cytokines, such as TNF-α and IL-6 and therefore targeting inflammation has been of therapeutic interest (George & Granger, 2011; Oyston et al., 2015; Kniotek & Boguska, 2017). Sildenafil might exert an anti-inflammatory response by inhibition of TNF-α and IL1β release and stimulation of IL-10 release (Ribaudo et al., 2016; Kniotek & Boguska, 2017). Indeed, prenatal sildenafil reduced TNF-α levels in maternal plasma and placenta in the rat model for preeclampsia and FGGR (Gillis et al., 2016) and reduced placental TNF-α and IL1β in the mice model for pregnancy loss. Administration of a different NO stimulating agent during healthy pregnancy lowered placental expression of IL1β and IL18 in sows (Luo et al., 2019). Our study showed only significant upregulation of the IL-10 signaling and pathway, but not TNF- α or IL1β. We speculate that most of the significantly upregulated gene sets promote protection to auto-immunity and innate immunity, which is necessary for embryonal implantation and placentation. This could potentially contribute to reducing pregnancy loss after prenatal sildenafil treatment when used earlier in pregnancy (Luna et al., 2015).

### Strengths and limitations

To our knowledge, this is the first study examining the effect of prenatal sildenafil administration during human pregnancies complicated by early-onset FGR on programming. One major strength is that we used two different types of tissue - with placenta representing maternal profile and HUVECs representing fetal profile - collected from a well-defined randomized controlled trial with severe early-onset FGR.

We acknowledge some limitations. This was a sub-study in which we collected samples from live births from the Dutch STRIDER trial and therefore does not fully represent the clin-

ical pregnancy outcomes. We attempted to minimize samples bias selection by using all the samples available, even if they were not paired. Because of the halt of the Dutch STRIDER study and because women were only recruited for this substudy in 2 of the 11 recruiting centers, this study is limited by a relatively low sample size. However, it also made the analysis of these samples unique and valuable. We used native HUVECs without prior selective culturing to be as close as possible to the in situ situation. Despite extensive washing, HUVEC samples might therefore have been contaminated with a few other fetal blood cells.

## Original purpose

The adverse *in utero* environment resulting in fetal growth restriction (FGR) predisposes the offspring to develop cardiorenal disease beyond the fetal developmental phase by altered epigenetic programming. Interest grew in prenatal administration of sildenafil after several animal studies showed improved fetal growth and long-term cardiovascular function.

An international STRIDER consortium emerged to evaluate the therapeutic potency of sildenafil to improve fetal growth in pregnancies complicated by FGR. In our sub-study (of the Dutch STRIDER trial) we aimed to study whether prenatal sildenafil influences developmental programming of cardiorenal health by examining whole genome RNA-sequencing and gene set enrichment analysis in human umbilical vein endothelial cells and placental tissue. We hypothesized that prenatal sildenafil alters fetal-placental programming profiles especially related to cardiorenal development and disease.

The interim-analysis of the Dutch STRIDER trial showed futility in combination with potentially increased mortality and morbidity in neonates. Hereafter all trials and inevitably our substudy were halted. This made our sub-study the first and last to investigate whether prenatal sildenafil administration in pregnancies complicated by FGR could influence fetal-placental programming of cardiorenal health.

Our study does not support the use of prenatal sildenafil for cardiorenal reprogramming. This is contrary to several animal studies, which may be suggestive of interspecies differences. Our results might be reassuring considering the negative clinical results of the STRIDER study and highlight the need for tissue collection from complicated pregnancies in humans to study developmental programming on a molecular or genetic level.

## Conclusion and future perspectives

FGR is associated with the developmental programming of cardiovascular and renal diseases later in life. Currently, no therapy exists to improve fetal growth or prevent these detrimental long-term consequences. Administration of PDE-5 inhibitors such as sildenafil during pregnancy showed beneficial effects on cardiovascular health in animal FGR offspring. However, our study in human pregnancies complicated by severe early-onset FGR did not show an effect of prenatal sildenafil administration on cardiovascular or renal programming. Future research is needed to understand whether an interspecies difference underlies these discrepancies or other differences in study design (such as dose) between animal and human studies. In order to progress, elucidation of (direct or indirect) underlying mechanisms and safety studies are of paramount importance in the evaluation of any new potential intervention.

## List of abbreviations

CAMERA, Correlation Adjusted MEan RAnk; CPM, count per million reads; EGM, endothelial cell growth medium; FDR, False Discovery Rates; FGR, fetal growth restriction; GA, gestational age; GSEA, gene set enrichment analysis; HELLP, Hemolysis, Elevated Liver enzymes and Low Platelet syndrome; HUVECs, human umbilical vein endothelial cells; IL1β, Interleukin 1β; MDS, Multidimensional scaling; $MgSO_4$, magnesium sulfate; NO, nitric oxide; PC, principle components; PCA, principal components analysis; PBS, phosphate-buffered saline; PPHN, persistent pulmonary hypertension; STRIDER, Sildenafil TheRapy In Dismal Prognosis Early-onset Fetal Growth Restriction; TMM, Trimmed mean of M-values; TNF-α, Tumor necrosis factor; VEGF, vascular endothelial growth factor.

## Ethics approval and consent to participate

The Dutch STRIDER trial (Clinical trial.gov identifier NCT02277132) was approved by the medical ethical committee of the AMC on 02-07-2014; protocol number 2014-131. The UMCU approved the study on 14-09-2015; protocol number 15-510/G-C. Prior to delivery, the STRIDER participants in this sub-study gave additional written informed consent for placental research (amendment approved on 29-01-2016, updated on 05-09-2017).

## Conflict of interest

None declared.

## Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to the Dutch privacy law to protect participants, but are partly and always coded available from the corresponding author on request. All data generated or analyzed during this study are included in this published article and its supplementary information files.

## References

Anderson, N. H., Sadler, L. C., McKinlay, C. J. D., & McCowan, L. M. E. (2016). INTERGROWTH-21st vs customized birthweight standards for identification of perinatal mortality and morbidity. *American Journal of Obstetrics and Gynecology, 214*(4), 509 1–509 7.

Barker, D. (2006). Adult consequences of fetal growth restriction. *Clinical Obstetrics and Gynecology, 49*, 270–83.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300.

Burton, G. J., & Jauniaux, E. (2018). Pathophysiology of placental-derived fetal growth restriction. *American Journal of Obstetrics and Gynecology, 218*(2), 745–761. https://doi.org/10.1016/j.ajog.2017.11.577

Chen, M., & Zhang, L. (2012). Epigenetic mechanisms in developmental programming of adult disease. Drugs Discovery Today, 16, 1007–1018. Demicheva, E., & Crispi, F. (2014). Long-term followup of intrauterine growth restriction: Cardiovascular disorders. *Fetal Diagnosis and Therapy, 36*(2), 143–153.

Dötsch, J., Alejandre-Alcazar, M., Janoschek, R., Nüsken, E., Weber, L. T., & Nüsken, K. D. (2016). Perinatal programming of renal function. *Current Opinion in Pediatrics, 28*(2), 188–194. https://doi.org/10.1097/MOP.0000000000000312

El Chami, H., & Hassoun, P. (2012). Immune and inflammatory mechanisms in pulmonary arterial hypertension. *Progress in Cardiovascular Diseases, 55*(2), 218–228. https://doi.org/10.1038/jid.2014.371

George, E. M., & Granger, J. P. (2011). Mechanisms and potential therapies for preeclampsia. *Current Hypertension Reports, 13*(4), 269–275.

George, E. M., Palei, A. C., Dent, E. A., & Granger, J. P. (2013). Sildenafil attenuates placental ischemiainduced hypertension. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology, 305*(4), 397–403.

Gillis, E. E., Mooney, J. N., Garrett, M. R., Granger, J. P., & Sasser, J. M. (2016). Sildenafil treatment ameliorates the maternal syndrome of preeclampsia and rescues fetal growth in the dahl salt-sensitive rat. *Hypertension, 67*(3), 647–653.

Gjerde, A., Lilla, B. S., Marti, H., Reisæter, A. V., & Vikse, B. E. (2020). Intrauterine growth restriction, preterm birth and risk of end-stage renal disease during the first 50 years of life. *Nephrology Dialysis Transplantation, 35*(7), 1157–1163.

Groom, K., McCowan, L. M., Mackay, L. K., Lee, A. C., Gardener, G., Unterscheider, J., Sekar, R., Dickinson, J. E., Muller, P., Reid, R. A., Watson, D., Welsh, A., Marlow, J., Walker, S. P., Hyett, J., Morris, J., Stone, P. R., & Baker, P. N. (2019). STRIDER NZAus: multicentre randomised controlled trial of sildenafil therapy in early-onset fetal growth restriction. *BJOG, 126*(8), 997–1006.

Hartman, R. J. G., Kapteijn, D. M. C., Haitjema, S., Bekker, M. N., Mokry, M., Pasterkamp, G., Civelek, M., & Ruijter, H. M. D. (2020). *Intrinsic transcriptomic sex differences in human endothelial cells at birth and in adults are associated with coronary artery disease targets*. Scientific Reports.

Henriksen, T., & Clausen, T. (2002). The fe-

tal origins hypothesis: Placental insufficiency and inheritance versus maternal malnutrition in wellnourished populations. *Acta Obstetricia et Gynecologica Scandinavica, 81*(2), 112–114.

Herrera, E. A., Cifuentes-Zúñiga, F., Figueroa, E., Villanueva, C., Hernández, C., Alegría, R., Arroyo-Jousse, V., Penaloza, E., Farias, M., Uauy, R., Casanello, P., & Krause, B. J. (2017). Nacetylcysteine, a glutathione precursor, reverts vascular dysfunction and endothelial epigenetic programming in intrauterine growth restricted guinea pigs. *The Journal of Physiology, 595*(4), 1077–1092.

Itani, N., Skeffington, K. L., Beck, C., & Giussani, D. A. (2017). Sildenafil therapy for fetal cardiovascular dysfunction during hypoxic development: Studies in the chick embryo. *Journal of Physiology, 595*(5), 1563–1573.

Kniotek, M., & Boguska, A. (2017). Sildenafil can affect innate and adaptive immune system in both experimental animals and patients. *Journal of Immunology Research*, 4541958.

Kobayashi, H., Yamataka, A., Okazaki, T., Lane, G. J., Puri, P., & Miyano, T. (2004). Increased levels of circulating adhesion molecules in neonates with congenital diaphragmatic hernia complicated by persistent pulmonary hypertension. *Pediatric Surgery International, 20*(1), 19–23. https://doi.org/10.1007/s00383-003-1072-8

Kooiman, J., Terstappen, F., van Wagensveld, L., Franx, A., Wever, K. E., Roseboom, T. J., Joles, J. A., Gremmels, H., & Lely, A. T. (2020). Conflicting effects of fetal growth restriction on blood pressure between human and rat offspring. *Hypertension, 75*(3), 806–818.

Krause, B. J., Hanson, M. A., & Casanello, P. (2011). Role of nitric oxide in placental vascular development and function. *Placenta, 32*(11), 797–805.

Kuebler, W. M., Bonnet, S., & Tabuchi, A. (2018). Inflammation and autoimmunity in pulmonary hypertension: Is there a role for endothelial adhesion molecules? *Pulmonary Circulation, 8*(2). https://doi.org/10.1177/2045893218757596

Liberzon, A., Birger, C., Ghandi, M., Jill, P., Tamayo, P., Jolla, L., & Jolla, L. (2016). *MSigDB H collection, 1*(6), 417–425.

Luna, R. L., Nunes, A. K. S., Oliveira, A. G. V., Araujo, S. M. R., Lemos, A. J. J. M., Rocha, S. W. S., Croy, B. A., & Peixoto, C. A. (2015). Sildenafil (viagra®) blocks inflammatory injury in LPS-induced mouse abortion: A potential prophylactic treatment against acute pregnancy loss? *Placenta, 36*(10), 1122–1129.

Luo, Z., Xu, X., Sho, T., Luo, W., Zhang, J., Xu, W., Yao, J., & Xu, J. (2019). Effects of n-acetylcysteine supplementation in late gestational diet on maternalplacental redox status, placental NLRP3 inflammasome, and fecal microbiota in sows. *Journal of Animal Science, 97*(4), 1757–1771.

Malhotra, A., Allison, B., Castillo-Melendez, M., Jenkin, G., Polglase, G., & Miller, S. (2019). Neonatal morbidities of fetal growth restriction: Pathophysiology and impact. *Frontiers in Endocrinology, 10*(55).

Man, A. W. C., Chen, M., Wu, Z., Reifenberg, G., Daiber, A., Münzel, T., Xia, N., & Li, H. (2020). Renal effects of fetal reprogramming with pentaerythritol tetranitrate in spontaneously hypertensive rats. *Frontiers in Pharmacology, 11*(April), 1–13.

Mills, V., Plows, J. F., Zhao, H., Oyston, C., Vickers, M. H., Baker, P. N., & Stanley, J. L. (2018). Effect of sildenafil citrate treatment in the eNOS knockout mouse model of fetal growth restriction on longterm cardiometabolic outcomes in male offspring. *Pharmacological Research, 137*(September), 122– 134.

Nüsken, E., Fink, G., Lechner, F., Voggel, J., Wohlfarth, M., Sprenger, L., Mehdiani, N., Weber, L. T., Liebau, M. C., Brachvogel, B., Dotsch, J., & Nüsken, K. D. (2020). Altered molecular signatures during kidney development after intrauterine growth restriction of different origins. *Journal of Molecular Medicine, 98*(3), 395–407. https:/ / doi.org/ 10.1007/ s00109-020-01875-1

Oyston, C. J., Stanley, J. L., & Baker, P. N. (2015). Potential targets for the treatment of preeclampsia. *Expert Opinion on Therapeutic Targets, 19*(11), 1517–1530.

Paauw, N. D., Terstappen, F., Ganzevoort, W., Joles, J. A., Gremmels, H., & Lely, A. T. (2017). Sildenafil during pregnancy: A preclinical meta-analysis on fetal growth and maternal blood pressure. *Hypertension, 70*(5), 998–1006.

Paauw, N. D., van Rijn, B. B., Lely, A. T., & Joles, J. A. (2016). Pregnancy as a critical window for blood pressure regulation in mother and child: Programming and reprogramming. *Acta Physiologica, 219*(1), 241–259.

Pels, A., Derks, J., Elvan-Taspinar, A., van Drongelen, J., de Boer, M., Duvekot, J., van Laar,

J., van yck, J., Al-Nasiry, S., Sueters, M., Post, M., Onland, W., van Wassenaer-Leemhuis, A., Naaktgeboren, C., Jakobsen, J. C., Gluud, C., Duijnhoven, R. G., Lely, T., Gordijn, S., & Group, D. S. T. R. I. D. E. R. T. (2020). Maternal sildenafil vs placebo for severe early-onset fetal growth restriction: A randomized clinical trial. *JAMA Network Open, 6*, 1–14.

Pels, A., Kenny, L. C., Alfirevic, Z., Baker, P. N., von Dadelszen, P., Gluud, C., Kariya, C. T., Mol, B. W., Papageorghiou, A. T., van Wassenaer-Leemhuis, A. G., Ganzevoort, W., Groom, K. M., & international STRIDER Consortium. (2017). STRIDER (sildenafil therapy in dismal prognosis early onset fetal growth restriction): An international consortium of randomised placebo-controlled trials. *BMC Pregnancy Childbirth, 17*(440), 1–8.

Rafikov, R., Nair, V., Sinari, S., Babu, H., Sullivan, J. C., Yuan, J. X. J., Desai, A. A., & Rafikova, O. (2019). Gender difference in damage-mediated signaling contributes to pulmonary arterial hypertension. *Antioxidants and Redox Signaling, 31*(13), 917–932. https://doi.org/10.1089/ars.2018.7664

Ribaudo, G., Angelo Pagano, M., Bova, S., & Zagotto, G. (2016). New therapeutic applications of phosphodiesterase 5 inhibitors (PDE5-is. *Current Medicinal Chemistry, 23*(12), 1239–1249.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). Edger: A bioconductor package for differential expression analysis of digital gene expression data. *26*(1), 139–40.

Sehgal, A., Alexander, B. T., Morrison, J. L., & South, A. M. (2020). Fetal growth restriction and hypertension in the offspring: Mechanistic links and therapeutic directions. *The Journal of Pediatrics, 224*, 115–123.

Sharp, A., Cornforth, C., Jackson, R., Harrold, J., Turner, M. A., Kenny, L. C., N., B. P., Johnstone, E. D., Khalil, A., von Dadelszen, P., Papageorghiou, A. T., Alfirevic, Z., & group, S. T. R. I. D. E. R. (2018). Maternal sildenafil for severe fetal growth restriction (STRIDER): A multicentre, randomised, placebocontrolled, double-blind trial. *The Lancet Child and Adolescent Health, 2*(2), 93–102.

Shue, E. H., Schecter, S. C., Gong, W., Etemadi, M., Johengen, M., Iqbal, C., Derderian, S. C., Oishi, P., Fineman, J. R., & Miniati, D. (2014). Antenatal maternally-administered phosphodiesterase type 5 inhibitors normalize eNOS expression in the fetal lamb model of congenital diaphragmatic hernia. *Journal of Pediatric Surgery, 49*(1), 39–45.

Sundrani, D., Roy, S., Jadhav, A., & Joshi, S. (2017). Sex-specific differences and developmental programming for diseases in later life. *Reproduction, F, 29*, 2085–2099.

Terstappen, F., Calis, J. J. A., Paauw, N. D., Joles, J. A., van Rijn, B. B., Mokry, M., Plosch, T., & Lely, A. T. (2020). Developmental programming in human umbilical cord vein endothelial cells following fetal growth restriction. *Clinical Epigenetics, 12*(1), 185.

Terstappen, F., Spradley, F. T., Bakrania, B. A., Clarke, S. M., Joles, J. A., Paauw, N. D., Garrett, M. R., Lely, A. T., & Sasser, J. M. (2019). Prenatal sildenafil therapy improves cardiovascular function in fetal growth restricted offspring of dahl salt-sensitive rats. *Hypertension, 73*(5), 1120–1127.

Turbeville, H. R., Johnson, A. C., Garrett, M. R., & Sasser, J. M. (2020). Sildenafil citrate does not reprogram risk of hypertension and chronic kidney disease in offspring of preeclamptic pregnancies in the dahl SS/jr rat. *Kidney360, 1*(6), 510–520. https://doi.org/10.34067/kid.0001062020

Von Dadelszen, P., Dwinnell, S., Magee, L., Carleton, B. C., Gruslin, A., Lee, B., Lim, K. I., Liston, R. M., Miller, S. P., Rurak, D., Sherlock, R. L., Skoll, M. A., Wareing, M. M., & Baker, P. N. (2011). Sildenafil citrate therapy for severe early-onset intrauterine growth restriction. *Research into Advanced Fetal Diagnosis and Therapy Group Baker, P. N*(5), 624–628.

Wareing, M., Myers, J. E., O'Hara, M., Kenny, L. C., Taggart, M. J., Skillern, L., Machin, I., & Baker, P. N. (2006). Phosphodiesterase-5 inhibitors and omental and placental small artery function in normal pregnancy and pre-eclampsia. *European Journal of Obstetrics, Gynecology, and Reproductive Biology, 127*(1), 41–49.

White, S. L., Perkovic, V., Cass, A., Chang, C. L., Poulter, N. R., Spector, T., Haysom, L., Craig, J. C., Al Salmi, I., Chadban, S. J., & Huxley, R. R. (2009). Is low birth weight an antecedent of CKD in later life? a systematic review of observational studies. *American Journal of Kidney Diseases, 54*(2), 248–261.

Wu, D., & Smyth, G. K. (2012). Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research, 40*(17), 133.

Wu, Z., Siuda, D., Xia, N., Reifenberg, G., Daiber, A., Munzel, T., Fostermann, U., & Li, H. (2015). Maternal treatment of spontaneously hypertensive rats with pentaerythritol tetranitrate reduces blood pressure in female offspring. *Hypertension, 65*(1), 232–237.

## Figure legends

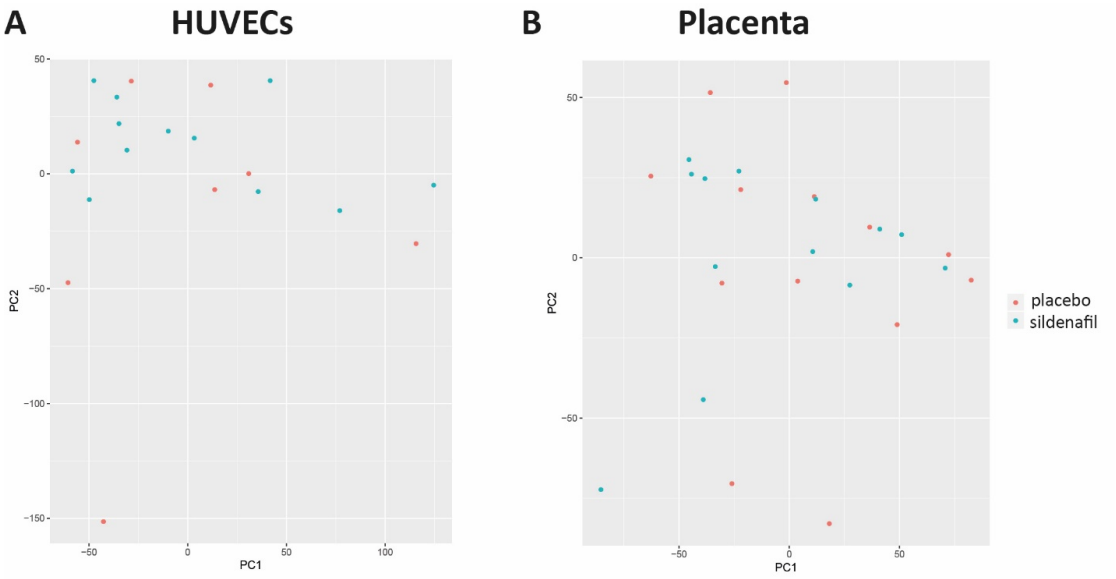



**Figure 1** Principal component analysis (PCA) plots of A) human umbilical vein endothelial cells (HUVECs) and of B) placenta.PC1 versus PC2 does not show clustering in treatment group sildenafil (blue) vs. placebo (red) in HUVECs or placental tissue.

## Tables

**Table 1** Maternal and neonatal characteristics

| | HUVECs | | | Placenta | | |
|---|---|---|---|---|---|---|
| | Sildenafil(n=12) | Placebo(n=8) | p-value | Sildenafil(n=13) | Placebo (n=13) | p-value |
| **Maternal characteristics during pregnancy** | | | | | | |
| Age, years | 35±6 | 31±3 | 0.19 | 34±6 | 33±6 | 0.66 |
| (pre-pregnancy) BMI, kg/m$^3$ | 23±3 | 24±6 | 0.74 | 24±4[#] | 26±7[#] | 0.29 |
| Preeclampsia/HELLP, % | 8 (67) | 3 (38) | 0.36 | 6 (46) | 5 (39) | 1.00 |
| Smoking, % | 1 (8) | 0 (0) | 1.00 | 1 (8) | 2 (15) | 1.00 |
| **Maternal medication during pregnancy** | | | | | | |
| Antihypertensive drugs, % | 8 (67) | 2 (25) | 0.17 | 6 (46) | 4 (31) | 0.69 |
| Antenatal steroids, % | 9 (75) | 7 (88) | 0.62 | 8 (62) | 7 (54) | 1.00 |
| $MgSO_4$, % | 2 (18)[#] | 0 (0) | 0.49 | 2 (15) | 0 (0) | 0.48 |
| GA start allocated drug, weeks | 25.0±2.0 | 24.5±2.2[#] | 0.64 | 24.0±2.5 | 25.0±2.4 | 0.33 |
| Duration allocated drug, days | 30.6±20.1 | 44.3±20.1[#] | 0.17 | 25.4±18.5 | 24.7±17.8 | 0.92 |
| **Delivery** | | | | | | |
| Caesarean section, % | 9 (75) | 6 (75) | 1.00 | 8 (62) | 6 (46) | 0.70 |
| Apgar at 5 min | 8 (3-10) | 8 (6-10) | 0.88 | 6 (0-9) | 8 (0-10) | 0.29 |
| **Neonatal characteristics** | | | | | | |
| Male gender, % | 7 (58) | 5 (63) | 1.00 | 8 (62) | 6 (46) | 0.70 |
| GA at birth, weeks | 30.8±4.3 | 32.2±3.7 | 0.45 | 27.8±1.9 | 29.4±4.1 | 0.21 |
| Birth weight, gram | 795 (430-2528) | 852 (580-2282) | 0.64 | 520(280-1005) | 770 (315-2385) | 0.23 |
| Birth weight, percentile | 3.6 (<0.01-16.7) | 1.0 (<0.01-4.7) | 0.22 | 0.1 (<0.01-13.3) | 0.7(<0.01-8.0) | 0.19 |
| *- <3$^{rd}$ percentile* | 6 (50) | 2 (25) | 0.37 | 8 (62) | 8 (62) | 1.00 |
| Survival, % | 8 (67) | 7 (88) | 0.60 | 5 (42) | 8 (62) | 0.43 |
| PPHN, % | 3 (25) | 0 (0) | 0.24 | 2 (15) | 0 (0) | 0.48 |

Data expressed as mean±SD, median (min-max), and n(%), which were respectively tested with independent t-test, Mann-Whitney, or Fisher exact. Magnesium sulfate ($MgSO_4$) was based on the maternal indication. [#] represents missing data of maximal one patient per group and therefore, the percentages are calculated based on the number of observations/measurements. GA, gestational age; HELLP, Hemolysis, Elevated Liver enzymes and Low Platelet syndrome; HUVECs, human umbilical vein endothelial cells; p, percentile; PPHN, persistent pulmonary hypertension.

**Table 2** Significantly different gene sets related to cardiovascular development or NO pathway between ***in vivo*sildenafil and placebo treated placental tissue samples**

| Gene set name | Up ordown | p-value | FDR | Brief description |
|---|---|---|---|---|
| GO_CELLULAR_RESPONSE_TO_VASCULAR_ENDOTHELIAL_GROWTH_FACTOR_STIMULUS | Up | 0.0013 | 0.0437 | Any process that results in a change in state or activity of a cell (movement, secretion, enzyme production, gene expression) as a result of a VEGF stimulus |
| GO_LEUKOCYTE_ADHESION_TO_VASCULAR_ENDOTHELIAL_CELL | Up | 0.0031 | 0.0848 | The attachment of a leukocyte to vascular endothelial cell via adhesion molecules |
| GO_NEGATIVE_REGULATION_OF_NITRIC_OXIDE_METABOLIC_PROCESS | Up | 0.0031 | 0.0848 | Any process that stops, prevents or reduces the frequency, rate or extent of nitric oxide metabolic process |
| BIOCARTA_AMI_PATHWAY | Up | 0.0036 | 0.0945 | Acute myocardial infarction is the condition of irreversible necrosis of the heart muscle that results from prolonged ischemia |

Ordered according to lowest false discovery rate (FDR). VEGF, vascular endothelial growth factor.

# In the era of whole transcriptome sequencing: Reflections on the Molecular Genetic Effect of Prenatal Sildenafil for Fetal Growth Restriction

Carsten F.J. Bakhuis[1], Marcel A.G. van der Heyden[2]

In this reflection article, we evaluate a sub study of the STRIDER trial by Terstappen et al., which investigated the molecular effects of prenatal sildenafil administration in pregnancies complicated by fetal growth restriction (FGR). Unfortunately, this trial revealed no clinical benefit and even an increased risk of persistent pulmonary hypertension in neonates. After an early trial cessation, this sub study tried to elucidate tissue-specific sildenafil effects by performing RNA sequencing on placental tissues and human umbilical vein endothelial cells. While no significant differences were found on gene level, modest pathway-level alterations (specifically in nitric oxide and immune signaling pathways) were observed. We here reflect on the methodological strengths of combining clinical and molecular data, but also point out limitations of this study such as the restricted gene set choice and the absence of an analysis stratified by neonatal outcome. For future drug repurposing studies, we highlight the importance of a broad molecular characterization of target tissues to fully explain effects that are observed in clinical trials.

**Keywords** *reflection, sildenafil, fetal growth restriction*

[1]The New Utrecht School, Utrecht, the Netherlands

[2]Department of Medical Physiology, University Medical Center Utrecht, the Netherlands

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
March 27, 2025
**Accepted**
April 28, 2025
**Published**
June 21, 2025

**Correspondence**
Department of Medical Physiology, University Medical Center Utrecht
M.A.G.vanderHeyden@umcutrecht.nl



## Introduction

Fetal growth restriction (FGR) is defined as the inability of a fetus to reach its intrinsic growth potential. The most used classification in clinical practice is of the American College of Obstetrics and Gynecology, which defines FGR as an estimated fetal weight below the tenth percentile (American College of Obstetrics and Gynecology, 2019; Nardozza et al., 2017). This occurs in 5-10% of all pregnancies in higher income countries (Damhuis et al., 2021; Frøen et al., 2004). FGR constitutes a major clinical challenge, as it is the second leading cause of perinatal mortality (Nardozza et al., 2017). This is especially true for early onset FGR (before 32 weeks of gestation), with a combined antenatal and neonatal mortality rate of 19% as presented in a recent systematic review (Pels, Beune, et al., 2020). Moreover, an association between FGR and diseases in adulthood has been observed, amongst others for cardiovascular and renal disease (Barker, 2006; Demicheva & Crispi, 2014; Frøen et al., 2004; White et al., 2009). The underlying etiologic mechanism of FGR is placental insufficiency, with or without maternal diseases, fetal chromosomal abnormalities, or an infection (Bruin et al., 2021).

Despite advancements in prenatal care, effective therapeutic interventions for FGR remain limited, necessitating continued research into potential treatments. One such candidate that has come to attention is sildenafil, a phosphodiesterase-5 inhibitor known for its vasodilatory effects, primarily used in the management of erectile dysfunction and pul-

monary hypertension (Pels et al., 2023). The hypothesis that sildenafil could enhance uteroplacental blood flow and thereby improve fetal growth was initially supported by promising results from various animal studies (Burke et al., 2016; Stanley et al., 2012), whilst some other preceding animal studies indicated a negative effect of sildenafil (Miller et al., 2009; Nassar et al., 2012). In 2017, results from a meta-analysis were published summarizing the results of several of these animal studies, indicating an overall potential beneficial effect of sildenafil administration (Paauw et al., 2017).

## The STRIDER consortium and the Dutch STRIDER trial

To evaluate the efficacy of sildenafil in human pregnancies complicated by FGR, the international Sildenafil TheRapy In Dismal prognosis Early-onset intrauterine growth Restriction (STRIDER) consortium was launched in 2014 (Pels et al., 2017). The STRIDER consortium consisted of research groups from several countries: New Zealand/Australia, Canada, the United Kingdom, and the Netherlands. Patients were included from 2014 to 2020 and results from individual studies were initially published separately (Groom et al., 2019; Paauw et al., 2017; Pels, Derks, et al., 2020; Sharp et al., 2018; von Dadelszen et al., 2022). The results from the Dutch STRIDER trial were published in 2020. In the Dutch trial, pregnant women with an amenorrhea duration between 20 and 30 weeks with severe FGR were randomized to either sildenafil or placebo treatment. However, in a planned interim analysis of the Dutch STRIDER trial, the hypothesized beneficial effect was not observed. Even so, a significantly increased risk of persistent pulmonary hypertension in neonates (PPHN) was observed. After this interim analysis, the further execution of the Dutch STRIDER trial was halted in 2018. The authors therefore describe their results based on all patients and samples included up until that moment (Pels, Derks, et al., 2020).

In an effort to explain the unexpectedly unfavorable observations of the trial and to further elucidate the effect of prenatal sildenafil administration on both the human fetus and placenta, a sub study was performed in the Dutch STRIDER trial. In this sub study, the effect of sildenafil on specifically cardiovascular and renal programming was studied using RNA sequencing (RNAseq; Terstappen et al., 2023). To discriminate between effects present within the fetal vascular system and in the placenta, both Human Umbilical Vein Endothelial Cells (HUVECs) samples and placental tissues of sildenafil-treated and placebo-treated patients underwent whole transcriptome RNAseq. In this reflection article, we will first go over some characteristics of RNAseq and HUVECs, summarize the study results, and then reflect thereon.

### Companion Article

Terstappen et al. (2023)

*Prenatal sildenafil and fetal-placental programming in human pregnancies complicated by fetal growth restriction: A retrospective gene expression analysis*

DOI: 10.36850/e16

## RNA sequencing and HUVECs: An introduction

RNA sequencing (RNAseq) is a powerful and increasingly used technique to analyze the complete transcriptome of a given sample. Using RNAseq, scientists can (amongst other things) evaluate the expression of certain genes by a quantification of the RNA levels of that specific gene. By comparing gene expression levels between specific patient groups, a so-called differential expression analysis can be performed (Khatoon et al., 2014). In the context of this study, RNAseq was used to compare the transcriptomic profiles from tissues derived from sildenafil-treated and placebo-treated patients, aiming to identify differentially expressed genes and certain pathways influenced by sildenafil.

This study used both placenta tissues and Human Umbilical Vein Endothelial Cells (HUVECs). HUVECs are a well-established model for studying vascular biology and endothelial cell function. The cells are derived from the endothelium of the umbilical vein after childbirth,

providing a representative cell type for examining vascular responses in vitro. However, that does cause some limitations with their usage. First, as HUVECs are obtained from the umbilical vein, they may not fully represent the total variety of fetal endothelial cells. Second, as they are harvested at birth, they may not completely reflect temporal changes which occurred during pregnancy and drug exposure. Also, as HUVECs are fetal tissue and not maternal tissue, sex differences between fetuses influence gene expression (Medina-Leyte et al., 2020). For that reason, the STRIDER sub study also corrected for this (Terstappen et al., 2023). Another limitation is that gene expression in HUVECs is generally heavily influenced by preterm birth (Medina-Leyte et al., 2020). As the mean gestational age at birth in the sildenafil treated group was approximately 2 weeks shorter than the placebo group (30.8 versus 32.2 weeks, respectively), this may have influenced the results from the gene expression analysis, although this difference in gestational age was not significant. Despite these limitations, HUVECs are in this setting the most accessible and relatively representative model to investigate the effect of sildenafil on fetal endothelial development.

### Summary of the results

In an effort to explain the potential effect of prenatal sildenafil administration on both the fetus and the placenta in the presence of FGR, the authors performed differential gene expression analysis within each tissue type (i.e., HUVEC or placenta) to compare patients with and without exposure of sildenafil.

First, they performed an overall differential gene expression analysis on all genes and emphasized specifically for genes known to be involved in cardiovascular or renal health or in the nitric oxide (NO) pathway. In this analysis, no significant differences were seen between the sildenafil- and placebo-treated groups. Second, the authors performed a so-called gene set analysis, where pathways consisting of specific genes are used as input and differences in pathway functioning can be explored. In this analysis, an upregulation of gene sets involved in the nitric oxide pathway and a varying up- and downregulation for immune pathway genes was observed in the placental tissues of the sildenafil-treated group. According to the authors, the upregulation of the three NO-related gene sets may represent the true pharmacological effect of sildenafil, although these differences were minimal and independent of usage duration of sildenafil. The observed upregulation of interleukin-10 related pathways might represent an anti-inflammatory response induced by the exposure to sildenafil, as the expression levels also correlated with the usage duration (Terstappen et al., 2023).

In general, the results of this study are therefore somewhat difficult to interpret. Although sildenafil may not directly influence cardiovascular or renal programming at the individual gene level, it may modulate broader biological pathways relevant to placental function and immune regulation. This therefore underscores the high complexity of developmental programming in pregnancy and the effects that FGR may have thereon.

In this reflection article, we aim to evaluate the approaches used in this sub study, explore alternative analytical strategies, and propose future directions for research.

## Reflection on "Prenatal sildenafil and fetal-placental programming in human pregnancies complicated by fetal growth restriction: A retrospective gene expression analysis"

The Dutch STRIDER trial added valuable information to the literature, as it sought to provide a treatment for a yet unmet clinical need. This contribution of an in-human clinical trial is especially valuable after the conflicting results of preceding animal studies, as these results may be difficult to translate to human physiology given the large interspecies differences that are often observed. Despite the negative results of the initial placebo-controlled trial, the authors still tried to elucidate the specific mechanisms of action of sildenafil in this sub study. While prenatal sildenafil administration does not improve pregnancy outcomes, this transcriptomic analysis provided valuable insights into potential molecular pathways that may be modulated by sildenafil exposure. Therefore, this study highlights the complexity of the pharmacological interactions that may occur when testing new interventions

to treat FGR, which is also relevant for future trials with other candidate treatments for this yet unmet clinical need.

As a general remark, we are very supportive of this type of translational research coupled to a clinical trial to better explain drug-tissue interactions observed in humans. In many cases, a clinical trial follows after an effect observation during in-vitro studies. However, considering a drug repurposing study such as the STRIDER trial, it is still critical to characterize the changes that certain tissues undergo when a drug is tested for a new application. In our opinion, this should be standard practice, regardless of an eventual beneficial or non-beneficial effect of the tested drug. Also, the broad approach of this study by using both HUVECs and placental tissues must be applauded. Our reflection will therefore mainly focus on potential other analyses which could have been performed with both the generated data in this study, and the samples themselves.

The authors themselves acknowledge their sample size as the main limitation, with usable samples of 13 sildenafil-treated patients (placental tissues, of whom 12 with HUVEC samples) and 13 placebo controls (placental tissue, of whom 8 with HUVEC samples; Terstappen et al., 2023). This hampered sample size was of course due to the early cessation of the Dutch STRIDER trial, but did limit the statistical power to detect more subtle molecular genetic alterations induced by the sildenafil treatment. This is especially true for the HUVEC group, where less samples were available as compared to the placental tissues. Nevertheless, one might expect that in this limited sample size the most relevant molecular genetic effect of sildenafil usage would have been detected as well. Therefore, this sub study did not yet provide us with a satisfactory answer to if and how sildenafil may influence the human placenta and/or the fetus. Still, the presence of such an influence may be logically deducted from the fact that, in the initial STRIDER trial, PPHN was observed significantly more frequent after sildenafil application (Pels, Derks, et al., 2020). Other processes or pathways, which have not been detected in this RNAseq study, may therefore be involved.

In the context of this sub-study, it must be noted that this transcriptomic approach to explain fetal and placental effects of sildenafil administration represents only a part of the potential complex processes involved. If we therefore want to better characterize the influence of sildenafil, other studies could focus on identifying other potential processes which may have been influenced. Examples of these studies could include DNA methylation analysis on the one hand, and proteomics (for example by means of mass spectrometry) to investigate post-translational changes on the other hand. As the authors of this sub study mention a potential application of sildenafil in early pregnancy to prevent pregnancy loss, it is advisable to first characterize the precise mechanisms of action and influence of this drug better.

Another possibility might have been a less restricted choice of gene sets for the differential gene expression analysis. Now, only gene sets known to be influential for cardiovascular or renal programming or to be involved in NO signaling were specifically studied. This approach does narrow the results, of course, where a potential effect of sildenafil through a primarily more unexpected (and therefore not included) pathway would go unnoticed. The authors did perform an overall analysis of differential gene expression, which did not show a clear distinction between the treated and untreated group. However, a targeted approach with gene sets of more pathways and processes may have shed even more light on the consequences of sildenafil application on the fetus and placenta.

In addition to this last remark, it may also be interesting to compare the HUVEC and placental samples of neonates with and without PPHN after sildenafil usage. Although this was not the primary goal of this study, it would have been a good subject for a follow-up study after the initial observation in the regular STRIDER trial of the significantly increased PPHN frequency. For example, within the sildenafil exposed group, differential gene expression analysis could have been performed to compare the three patients with PPHN to the patients without PPHN. Despite the very limited sample size of such an analysis, this may have given an indication of the presence of critical sildenafil-influenced pathways, and the respective relevant gene sets could then have been studied in the overall trial population. Ultimately, such a comparison may also help to identify patients which are at risk for PPHN in the future if silde-

nafil is to be applied during pregnancy for other potential indications.

In conclusion, despite the effort of the STRIDER consortium, sildenafil did not prove to be the "golden bullet" for the treatment of FGR. Although the authors have done extensive research to gain insights on a molecular-genetic level of the influence of sildenafil, this study did not yet provide us with a satisfactory answer to explain the observed effects in the trial. To reduce pregnancy losses or adverse neonatal outcomes due to FGR, future research should focus on an even better characterization of the mechanisms involved in FGR. By doing so, the scientific community can come up with the best suitable targets or even with candidate drugs for this yet unmet clinical need.

## References

American College of Obstetricians and Gynecologists' Committee on Practice Bulletins—Obstetrics and the Society for Maternal-Fetal Medicine. (2019). ACOG Practice Bulletin No. 204: Fetal Growth Restriction. *Obstetrics and Gynecology, 133*(2), e97–e109. https://doi.org/10.1097/AOG.0000000000003070

Barker, D. J. (2006). Adult consequences of fetal growth restriction. *Clinical Obstetrics and Gynecology, 49*(2), 270–283. https://doi.org/10.1097/00003081-200606000-00009

Bruin, C., Damhuis, S., Gordijn, S., & Ganzevoort, W. (2021). Evaluation and management of suspected fetal growth restriction. *Obstetrics and Gynecology Clinics of North America, 48*(2), 371–385. https://doi.org/10.1016/j.ogc.2021.02.007

Burke, S. D., Zsengellér, Z. K., Khankin, E. V., Lo, A. S., Rajakumar, A., DuPont, J. J., McCurley, A., Moss, M. E., Zhang, D., Clark, C. D., Wang, A., Seely, E. W., Kang, P. M., Stillman, I. E., Jaffe, I. Z., & Karumanchi, S. A. (2016). Soluble fms-like tyrosine kinase 1 promotes angiotensin II sensitivity in preeclampsia. *The Journal of Clinical Investigation, 126*(7), 2561–2574. https://doi.org/10.1172/JCI83918

Damhuis, S. E., Ganzevoort, W., & Gordijn, S. J. (2021). Abnormal fetal growth: Small for gestational age, fetal growth restriction, large for gestational age: Definitions and Epidemiology. *Obstetrics and Gynecology Clinics of North America*, *48*(2), 267–279. https://doi.org/10.1016/j.ogc.2021.02.002

Demicheva, E., & Crispi, F. (2014). Long-term follow-up of intrauterine growth restriction: Cardiovascular disorders. *Fetal Diagnosis and Therapy, 36*(2), 143–153. https://doi.org/10.1159/000353633

Frøen, J. F., Gardosi, J. O., Thurmann, A., Francis, A., & Stray-Pedersen, B. (2004). Restricted fetal growth in sudden intrauterine unexplained death. *Acta Obstetricia et Gynecologica Scandinavica, 83*(9), 801–807. https://doi.org/10.1111/j.0001-6349.2004.00602.x

Groom, K. M., McCowan, L. M., Mackay, L. K., Lee, A. C., Gardener, G., Unterscheider, J., Sekar, R., Dickinson, J. E., Muller, P., Reid, R. A., Watson, D., Welsh, A., Marlow, J., Walker, S. P., Hyett, J., Morris, J., Stone, P. R., & Baker, P. N. (2019). STRIDER NZAus: A multicentre randomised controlled trial of sildenafil therapy in early-onset fetal growth restriction. *BJOG: An International Journal of Obstetrics and Gynaecology, 126*(8), 997–1006. https://doi.org/10.1111/1471-0528.15658

Khatoon, Z., Figler, B., Zhang, H., & Cheng, F. (2014). Introduction to RNA-Seq and its applications to drug discovery and development. *Drug Development Research, 75*(5), 324–330. https://doi.org/10.1002/ddr.21215

Medina-Leyte, D. J., Domínguez-Pérez, M., Mercado, I., Villarreal-Molina, M. T., & Jacobo-Albavera, L. (2020). Use of Human Umbilical Vein Endothelial Cells (HUVEC) as a model to study cardiovascular disease: A review. *Applied Sciences, 10*(3), Article 938. https://doi.org/10.3390/app10030938

Miller, S. L., Loose, J. M., Jenkin, G., & Wallace, E. M. (2009). The effects of sildenafil citrate (Viagra) on uterine blood flow and well being in the intrauterine growth-restricted fetus. *American Journal of Obstetrics and Gynecology, 200*(1), 102.e1–102.e1027. https://doi.org/10.1016/j.ajog.2008.08.029

Nardozza, L. M., Caetano, A. C., Zamarian, A. C., Mazzola, J. B., Silva, C. P., Marçal, V. M., Lobo, T. F., Peixoto, A. B., & Araujo Júnior, E. (2017). Fetal growth restriction: Current knowledge. *Archives of Gynecology and Obstetrics, 295*(5), 1061–1077. https://doi.org/10.1007/s00404-017-4341-9

Nassar, A. H., Masrouha, K. Z., Itani, H., Nader, K. A., & Usta, I. M. (2012). Effects of sildenafil in Nω-nitro-L-arginine methyl ester-induced intrauterine growth restriction in a rat model. *American Journal of Perinatology, 29*(6),

429–434. https:/ / doi.org/ 10.1055/ s-0032-1304823

Paauw, N. D., Terstappen, F., Ganzevoort, W., Joles, J. A., Gremmels, H., & Lely, A. T. (2017). Sildenafil during pregnancy: A preclinical meta-analysis on fetal growth and maternal blood pressure. *Hypertension, 70*(5), 998–1006. https://doi.org/10.1161/HYPERTENSIONAHA.117.09690

Pels, A., Kenny, L. C., Alfirevic, Z., Baker, P. N., von Dadelszen, P., Gluud, C., Kariya, C. T., Mol, B. W., Papageorghiou, A. T., van Wassenaer-Leemhuis, A. G., Ganzevoort, W., Groom, K. M., & International STRIDER Consortium. (2017). STRIDER (Sildenafil TheRapy In Dismal prognosis Early onset fetal growth Restriction): An international consortium of randomised placebo-controlled trials. *BMC Pregnancy and Childbirth, 17*(1), Article 440. https://doi.org/10.1186/s12884-017-1594-z

Pels, A., Beune, I. M., van Wassenaer-Leemhuis, A. G., Limpens, J., & Ganzevoort, W. (2020). Early-onset fetal growth restriction: A systematic review on mortality and morbidity. *Acta Obstetricia et Gynecologica Scandinavica*, *99*(2), 153–166. https://doi.org/10.1111/aogs.13702

Pels, A., Derks, J., Elvan-Taspinar, A., van Drongelen, J., de Boer, M., Duvekot, H., van Laar, J., van Eyck, J., Al-Nasiry, S., Sueters, M., Post, M., Onland, W., van Wassenaer-Leemhuis, A., Naaktgeboren, C., Jakobsen, J. C., Gluud, C., Duijnhoven, R. G., Lely, T., Gordijn, S., ..., & the Dutch STRIDER Trial Group. (2020). Maternal sildenafil vs placebo in pregnant women with severe early-onset fetal growth restriction: A randomized clinical trial. *JAMA Network Open, 3*(6), Article e205323. https://doi.org/10.1001/jamanetworkopen.2020.5323

Pels, A., Ganzevoort, W., Kenny, L. C., Baker, P. N., von Dadelszen, P., Gluud, C., Kariya, C. T., Leemhuis, A. G., Groom, K. M., Sharp, A. N., Magee, L. A., Jakobsen, J. C., Mol, B. W. J., & Papageorghiou, A. T. (2023). Interventions affecting the nitric oxide pathway versus placebo or no therapy for fetal growth restriction in pregnancy. *The Cochrane Database of Systematic Reviews, 7*(7), Article CD014498. https://doi.org/10.1002/14651858.CD014498

Sharp, A., Cornforth, C., Jackson, R., Harrold, J., Turner, M. A., Kenny, L. C., Baker, P. N., Johnstone, E. D., Khalil, A., von Dadelszen, P., Papageorghiou, A. T., Alfirevic, Z., & STRIDER group. (2018). Maternal sildenafil for severe fetal growth restriction (STRIDER): A multicentre, randomised, placebo-controlled, double-blind trial. *The Lancet: Child & Adolescent Health, 2*(2), 93–102. https:/ / doi.org/ 10.1016/ S2352-4642(17)30173-6

Stanley, J. L., Andersson, I. J., Poudel, R., Rueda-Clausen, C. F., Sibley, C. P., Davidge, S. T., & Baker, P. N. (2012). Sildenafil citrate rescues fetal growth in the catechol-O-methyl transferase knockout mouse model. *Hypertension, 59*(5), 1021–1028. https://doi.org/10.1161/HYPERTENSIONAHA.111.186270

Terstappen, F., Plösch, T., Calis, J. J. A., Ganzevoort, W., Pels, A., Paauw, N. D., Gordijn, S. J., van Rijn, B. B., Mokry, M., & Lely, A. T. (2023). Prenatal sildenafil and fetal-placental programming in human pregnancies complicated by fetal growth restriction: A retrospective gene expression analysis [Special issue]. *Journal of Trial and Error*. https://doi.org/10.36850/e16

von Dadelszen, P., Audibert, F., Bujold, E., Bone, J. N., Sandhu, A., Li, J., Kariya, C., Chung, Y., Lee, T., Au, K., Skoll, M. A., Vidler, M., Magee, L. A., Piedboeuf, B., Baker, P. N., Lalji, S., & Lim, K. I. (2022). Halting the Canadian STRIDER randomised controlled trial of sildenafil for severe, early-onset fetal growth restriction: Ethical, methodological, and pragmatic considerations. *BMC Research Notes, 15*(1), Article 244. https://doi.org/10.1186/s13104-022-06107-y

White, S. L., Perkovic, V., Cass, A., Chang, C. L., Poulter, N. R., Spector, T., Haysom, L., Craig, J. C., Salmi, I. A., Chadban, S. J., & Huxley, R. R. (2009). Is low birth weight an antecedent of CKD in later life? A systematic review of observational studies. *American Journal of Kidney Diseases, 54*(2), 248–261. https://doi.org/10.1053/j.ajkd.2008.12.042

# Partial Endothelial Trepanation versus Deep Anterior Lamellar Keratoplasty in keratoconus patients: Results of the PENTACON trial

Robert P.L. Wisse[1], Cathrien A. Eggink[2], Bart T.H. van Dooren[3], Allegonda van der Lelij[1]

The purpose of this research was to report on the surgical safety and outcomes of two distinct techniques of anterior lamellar corneal surgery: the Big Bubble Deep Anterior Lamellar Keratoplasty (DALK) versus Busin's Partial ENdothelial Trepanation (PET) in addition to anterior lamellar keratoplasty in keratoconus patients. In short, the PENTACON trial. In this multicenter trial, patients were randomized to receive either a DALK or PET procedure. Primary outcome was the occurrence of a intra-operative complication necessitating conversion to a full-thickness corneal graft. Secondary outcomes were uncorrected and best corrected visual acuity (UCVA/BCVA), manifest refraction, corneal astigmatism, and adverse events at 1 year follow-up. Fourteen eyes of 14 patients were enrolled in this study. At baseline, mean logMAR UCVA and BCVA were 1.59 (*SD* = 0.35) and 0.89 (*SD* = 0.69) respectively. Mean thinnest pachymetry was 322 (*SD* = 66µm), with a mean Kmax of 76.7 (*SD* = 14.1D). In five of 13 surgeries a full-thickness conversion occurred (DALK:PET 3:2, (*p* = .592). Overall, logMAR UCVA and BCVA increased to 0.52 (*SD* = 0.20, (*p* = .003) and 0.26 (*SD* = 0.36, (*p* = .03) respectively at 1 year follow-up. Mean refractive astigmatism was 3.8 (*SD* = 2.2D). No significant differences were observed between both treatment groups for any of the secondary outcomes parameters. No conclusions can be drawn on the primary outcome based on this underpowered clinical trial. However, the PET technique was not as safe as expected. A low trial inclusion rate and lack of scientific equipoise prompted trial termination. We regard it our ethical obligation to report these results.

**Keywords** *keratoconus, lamellar surgery, surgical safety, clinical trial, DALK*

[1] University Medical Center, Utrecht, the Netherlands
[2] University Medical Center St. Radboud, Nijmegen, the Netherlands
[3] Amphia Ziekenhuis, Breda, the Netherlands

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
June 25, 2024
**Accepted**
March 18, 2025
**Published**
May 25, 2025

**Correspondence**
University Medical Center Utrecht
Rplwisse@gmail.com



## Introduction

Keratoconus is a progressive corneal condition characterized by irregular refractive properties that reduce visual acuity. Keratoconus usually arises in adolescence, is bilateral, and has an estimated incidence of 1:2,000 (Kennedy et al., 1986). Treatment is aimed at improving vision, principally using (rigid) gas permeable contact lenses. With progression of the disease, non-correctable refractive abnormalities and/or corneal scars arise. For these advanced stages of keratoconus, and in contact lens intolerance, a corneal transplant is a viable treatment modality.

The first corneal transplant for keratoconus was conducted in 1936 by Ramon Castroviejo in New York's Columbia Presbyterian Medical Centre. Ever since, corneal grafting has been subject to many technical developments. With the advent of refractive surgery in the 1990s (Buratto et al., 1992), equipment was developed to split a cornea in horizontal lamellae. This made partial thickness grafting possible, tailoring grafts according to the nature and location of corneal pathology. For keratoconus,

## Take-home message

The PENTACON trial, which compared Big Bubble Deep Anterior Lamellar Keratoplasty (DALK) techniques in advanced keratoconus patients, highlighted major challenges in clinical research. These included identifying feasible patient selection criteria, managing a prolonged study, and coordinating a multicenter design. Although the trial was underpowered and ultimately terminated, it underscores the ethical imperative to report all findings. These techniques have a steep learning curve that is frequently underreported.

only the affected anterior part of the cornea needs to be transplanted: the Deep Anterior Lamellar Keratoplasty (DALK). This technique circumvents transplanting the host endothelium, leading to a lower rate of graft rejection (Ang et al., 2008; Fontana et al., 2007). The main drawback of DALK is the risk of accidental corneal perforation during surgery, as the fragile Descemet membrane can easily rupture, potentially requiring conversion to a full-thickness graft. To prevent inadvertent perforation, several techniques are described to split the stroma from the posterior lying Descemet membrane and corneal endothelium, using either fluid (Amayem & Anwar, 2000), viscoelastic devices (Melles et al., 2000), or air (Anwar & Teichmann, 2002). Failure and perforation are described in 20% of cases, though, leading to poor surgical predictability (Cheng et al., 2011). The DALK techniques require a long learning curve, and the reported perforation rates might be an underestimate (Kasbekar et al., 2014).

To circumvent this problem, a technique was developed in which, in addition to a mechanized anterior lamellar keratoplasty, a Partial Endothelial Trepanation (PET) is performed. This technique was first performed by Prof. Massimo Busin, in the Villa Serena Hospital, Forli, Italy (Busin et al., 2012). The endothelium and Descemet membrane are paracentrally and circularly loosened, but a certain proportion is left intact. This 'island' is able to mold to the healthy donor curvature. By doing this, the surgeon can retain a safer graft thickness margin leading to a lowered number of preoperative perforations. The introduction of PET is believed to make corneal grafting safer and more predictable.

Here, we study the outcomes of this new technique in a randomized clinical trial, with the DALK technique as comparator technique: the Partial Endothelial Trepanation in addition to anterior lamellar keratoplasty in keratoconus patients, or the PENTACON, trial. Our primary goal was to assess the surgical safety of both techniques. Secondly, we assessed secondary treatment outcomes in terms of visual acuity, manifest refraction, corneal astigmatism, and endothelial cell density at 1 year post-treatment.

## Materials and methods

### Study design

This multicenter randomized clinical trial was conducted from March 2011 until June 2015. Study participation was granted by the University Medical Center Utrecht, University Medical Center Nijmegen St. Radboud, Rotterdam Eye Hospital, Amphia Ziekenhuis Breda, and Westfriesgasthuis Hoorn. The conduction of this study was approved by the Ethics Review Board of all participating centers and was performed in accordance with local laws, the European guidelines of Good Clinical Practice, and the tenets of the Declaration of Helsinki. The study was registered at ISCRTN (no° ISRCTN39068025) and clinicaltrials.gov (no° 30756.041.10).

Patients eligible for study participation were randomized using a permutated block size and were stratified for the presence of atopic diseases. The web-based randomization tool was hosted by our institutions biostatistical department (UMCU Julius Center).

### Patient selection

Inclusion criteria included: age ≥ 18 years, clinical and topographic evidence of keratoconus (KISA% index >100%), and reduced best corrected visual acuity from corneal scarring or contact lens intolerance (Rabinowitz & Rasheed, 1999; Tang et al., 2005).

Exclusion criteria included: prior corneal or refractive surgery, corneal thickness <300 µm, corneal steepness preventing suction ring

placement, corneal endothelial disease on specular microscopy, or any other significant ocular pathology that could reduce visual acuity beyond keratoconus itself.

### Primary and secondary outcomes

The event of a surgical complication necessitating conversion to a full-thickness corneal graft was considered as primary outcome parameter. Hereto, all surgical and post-operative adverse events and protocol deviations were recorded in study specific case report forms.

Secondary study objectives focused on the effectiveness of both techniques at 6 months and 1 year follow-up: uncorrected and best spectacle corrected visual acuity (UCVA/BCVA), manifest refraction, corneal astigmatism, contact lens use (soft/rigid/scleral) or spectacle use, graft rejection and failure rate, corneal endothelial function, and correlation of outcomes with atopic constitution. Graft rejection was assessed by slit lamp examination (Folks, 2005). Graft failure is related to endothelial cell dysfunction and graded concordantly as corneal endothelial disease. Atopical constitution is defined by the presence of allergic conjunctivitis at time of screening or confirmation of atopy (e.g. allergy, asthma, eczema, laboratory testing with elevated IgE levels) by patient history. All patients were routinely screened for total IgE serum levels.

### Clinical protocol and used equipment

Examinations were scheduled at baseline, at 6, and at 12 months follow up. The ophthalmic examination consisted of a brief history, use of (ocular) medication, use of visual aids (spectacles/contact lenses), and the occurrence of adverse events. UCVA and BCVA were assessed using an EDTRS visual acuity chart. Manifest refraction was taken by an optometrist or ophthalmic assistant. Slitlamp examination focused on the presence of corneal pathology, corneal clarity, and suture related complications. Hereto, dedicated case report forms were employed. A dilated fundus exam assessed the incidence of cataract, glaucoma, or macular disease.

Corneal topography and pachymetry were acquired using the Oculus Pentacam HR Type 70900 (Oculus Optikgeräte GmbH, Wetzlar, Germany). Endothelial cell counts were acquired with the Topcon Sp-3000p, Topcon Corporation, Tokyo, Japan. Intraocular pressure was measured using the Topcon CT-80. If unattainable, a Goldmann applanation tonometer was used.

### Surgical technique and donor preparation

All donor corneas were supplied by the European Cornea Bank, Beverwijk, conform EEBA medical standards (European Eye Bank Association, 2008).

Patients were randomly divided into two groups, which received different treatments. Group A received a Partial Endothelial Trepanation (PET) in addition to anterior lamellar keratoplasty Part of the described technique is published by Busin (Busin et al., 2012). We applied the technique following these instructions: The donor cornea is mounted on an artificial anterior chamber (ALTK, Moria S.A., Antony, France) with the epithelium up, and an anterior corneal lamella is cut with a 350µm microkeratome head and a hand-driven microkeratome (CBm, Moria S.A., Antony, France). Then the anterior corneal lamella of the recipient is prepared by applying the suction ring to the eye of the patient, and the intraocular pressure is increased to >65 mmHg. Balanced salt solution (ALCON, Fort Worth, Texas, USA) is instilled on the corneal surface and the same hand-driven microkeratome is advanced in the tract until the anterior lamella is completely severed from the underlying recipient stroma. At least 100µm residual tissue should be left in place. Thereafter a partial trepanation with a 6.5 mm disposable hand trephine is made into the remaining stroma. In this grove of the remaining tissue, including Descemet's membrane and endothelium, a cut is completed manually and oblique with a Thornton knife over 180-270°. This small 'island' will stay in place. The diameter of the exposed stromal bed is measured with a caliper, and the diameter of the donor graft is chosen accordingly. Finally, the lamellar graft is sutured in place under tension of 16 interrupted 10-0 nylon sutures. After removal of the speculum, the eye is patched.

For Group B we applied the conventional Deep Anterior Lamellar Keratoplasty (DALK)

type Big Bubble technique according to Anwar and Teichmann (2002).

**Table 1** Baseline characteristics of both treatment groups

| | | PET | DALK | *p** |
|---|---|---|---|---|
| Gender (% male) | | 86% | 33% | 0.06 |
| Age | | 36.4 ±10.8 | 40.5 ±14.2 | 0.56 |
| Atopy | | 57% | 67% | 0.97 |
| UCVA (logMAR) | | 1.76 ±0.21 | 1.36 ± 0.39 | 0.14 |
| BCVA (logMAR) | | 1.12 ±0.79 | 0.74 ±0.65 | 0.42 |
| Manifest refraction | | | | |
| Sphere (D) | -9.25 ±6.33 | -4.17 ±5.41 | 0.21 | |
| Cylinder (D) | -2.88 ±2.22 | -3.71 ±1.96 | 0.55 | |
| Kflat (D) | | 58.65 ±6.25 | 58.45 ±6.69 | 0.96 |
| Ksteep (D) | | 66.45 ±10.12 | 63.23 ±7.00 | 0.54 |
| Kmax (D) | | 78.05 ±17.43 | 75.37 ±11.24 | 0.76 |
| Corneal astigmatism (D) | | 3.32 ±1.99 | 4.77 ±3.08 | 0.36 |

PET: Partial Endothelial Trepanation. DALK: Deep Anterior Lamellar Keratoplasty. UCVA: uncorrected visual acuity. BCVA: best corrected visual acuity. D: diopter. *independent students t-test

### Statistical analysis and power analysis

Baseline measurements between the treatment groups were compared using an independent samples t-test. Fischer's exact test (two-tailed) was used to determine the relation between treatment and risk of conversion to a perforating keratoplasty. Decimal visual acuity was converted to the logarithm of the minimal angle of resolution (logMAR). Normality and homoscedasticity of the residuals were tested visually and in a Q-Q plot and scatterplot, respectively. A $p$-value < .05 was considered statistically significant. Data are recorded as mean ± standard deviation. All tests were performed in SPSS version 22.0 for Windows.

With an expected perforation risk reduction of 85% (current DALK ratio = 20%, 3% perforation reported by Busin et al. (2012)), incorporating a sequential power calculation with a two-sided alpha 0.05 and beta 0.80, approximately 30 patients needed to be included in each treatment arm (Chow et al., 2003; D'Agostino et al., 1988).

## Results

### Clinical characteristics

A total of 14 eyes from 14 patients were enrolled in this trial. Two external centers participated (Radboud UMC $n$ = 2, Amphia Ziekenhuis Breda $n$ = 1).One patient postponed his surgery after randomization and was excluded from analysis. Six DALK procedures and 7 PET procedures were therefore included. Two randomized cases (one DALK, one PET) developed a corneal hydrops while on the waiting list for surgery, and following the intention-to-treat analysis both were included. Mean age was 38.3 years ($SD$ = 12.1) and 61.5% of the patients was male. At baseline, mean logMAR UCVA and BCVA were 1.59 ($SD$ = 0.35) and 0.89 ($SD$ = 0.69) respectively. Mean refractive astigmatism was 3.4 ($SD$ = 2.0D), mean IOP 10 ($SD$ = 2.1mmHg), and mean thinnest pachymetry 322 ($SD$ = 66µm). Endothelial cell counts were only attainable in two cases (2110 and 2558 cells/mm$^2$). Topographic indices on average were a $K_{max}$ of 76.7 ($SD$ = 14.1D), $K_{flat}$ 58.6 ($SD$ = 6.17D), $K_{steep}$ 64.8 ($SD$ = 8.5D), and an astigmatism of 4.0 ($SD$ = 2.6D). Baseline characteristics did not differ significantly between both groups (see table 1).

### Donor characteristics

All patients received their donors from the European Cornea Bank, Beverwijk, the Netherlands. Mean donor age was 61.3 years ($SD$ = 9.8) and the average time between death and enucleation was 15½ hours. On average the difference between the age of the donor and patient was 21.4 years ($SD$ = 12.2), range 4-42. Mean ECD was 2692 ($SD$ = 170 cells/mm$^2$), range 2400-3000. Donor characteristics did not differ significantly between both groups (data not shown).

### Primary outcome

The primary study outcome was defined as the incidence of surgical adverse events necessitating conversion to a penetrating keratoplasty. Adverse events occurred in ten of thirteen surgeries. Five of thirteen surgeries were converted to perforating keratoplasties (DALK:PET 3:2, $p$ = 0.592, Fisher's exact test),

including both cases with a previous corneal hydrops. Only two surgeries (both PET) had no complications. Notable protocol deviations included five full perforations, two microperforations (both DALK), three poor microkeratome cuts (all PET), three post-operative rebubblings (DALK:PET 1:2), and two pre-Descemetic preparations (both DALK).

## Original purpose

The PENTACON trial was designed to evaluate the surgical safety and clinical outcomes of two advanced corneal transplantation techniques—Big Bubble Deep Anterior Lamellar Keratoplasty (DALK) and Busin's Partial Endothelial Trepanation (PET) in patients with keratoconus. Keratoconus is a progressive corneal disease characterized by thinning and protrusion, leading to significant visual impairment. Traditional treatment approaches, including full-thickness corneal transplantation, carry a risk of graft rejection and complications, and the DALK technique is know for its steep learning curve and relatively high rate of intra- and post-operative complications. PET was conceptualized as a technically less demanding technique, potentially yielding comparable outcomes.

Our primary goal was therefore to assess whether the PET technique could offer a safer alternative to the DALK procedure. Secondary objectives included evaluating the visual acuity outcomes, refractive stability, corneal astigmatism, and adverse events 1 year post-operatively in both treatment arms. By comparing these parameters, we aimed to determine if the PET technique could provide comparable or superior visual rehabilitation while enhancing surgical safety and predictability.

In essence, the PENTACON trial was an endeavor to innovate and refine corneal transplantation techniques to improve patient outcomes, reduce surgical complications, and ultimately enhance the quality of life for individuals affected by keratoconus. Through rigorous multicenter collaboration and randomized controlled trial methodology, we aimed to contribute substantial evidence to guide clinical practice in corneal surgery.

### Secondary outcomes

Secondary outcomes were assessed at 6 months and 12 months post-operatively. Overall, at 6 months mean logMAR UCVA and BCVA increased to 0.93 (*SD* = 0.27, *p* = .02) and 0.48 (*SD* = 0.27, *p* = .15) respectively. At 12 months this further increased to 0.52 (*SD* = 0.20, *p* = .003) and 0.26 (*SD* = 0.36, *p* = .03). The following parameters are only reported at the 12 months assessment since topographic data and manifest refraction were often not attainable at the 6 months' time point. Due to the low number of cases only overall outcomes were reported; a valid comparison between both techniques was not feasible. All corneas were clear (one or two out of six) at the final follow-up visit, and all sutures were removed. Two PET cases had some Descemet folds, however (Snellen BCVA 0.45 & 0.55). One case (DALK) was suspected of an epithelial rejection and treated subsequently (Snellen BCVA 0.7). Two-thirds of the patients used scleral contact lenses after their surgery. Endothelial cell densities were too often unattainable or not recorded; only three viable measurements were recorded, data not shown. Mean refractive astigmatism was 3.8 (*SD* = 2.2D), with one case (DALK) of 8D astigmatism. On average the topographical indices were a $K_{max}$ of 53.0 (*SD* = 2.8D), $K_{flat}$ 42.4 (*SD* = 5.1), $K_{steep}$ 46.5 (*SD* = 3.5D), and an astigmatism of 4.9 (*SD* = 3.4D). No significant differences were observed between both treatment groups for any of the secondary outcomes parameters. No long term sequelae like suture related complications, graft failure, cataract, glaucoma, or ocular hypertension were noted during trial follow-up.

## Discussion

In general, no solid conclusions can be drawn with regards to the primary outcome based on this underpowered clinical trial. Whether the partial trepanation technique proposed by Busin is superior to the regular DALK technique in terms of surgical safety is still open for debate. On average, UCVA and BCVA improved significantly after 12 months, and visual acuity improved in all eyes. Post-operative mean refractive and topographic astigmatism were in line with other studies (Cheng et al., 2011; Söğütlü Sari et al., 2012). No long term se-

quelae from corneal surgery were recorded, although two-thirds of the patients used (scleral) contact lenses at the 12 months follow-up. Some findings of this study however deserve to be discussed.

Firstly, this trial was heavily underpowered. Consistent with Tolstoy's (1877/1997, p. 1) reference to (un)happy families, numerous disruptive events arose during the study. The most significant issues were:

1. challenges integrating the trial into routine clinical practice,
2. the introduction of corneal crosslinking mid-study (Godefrooij et al., 2016), and
3. a narrow surgical indication (mild cases do well with contact lenses, and severe cases often have post-hydrops scarring that makes them unsuitable).

During the 4 years that the trial was open for participation only 14 eyes were included, and we considered it unrealistic that the predefined power of 60 inclusions could eventually be met. Finally, we lost scientific equipoise: We could no longer maintain the belief that both treatments were equally advisable for our patients. We considered the PET a safer alternative to the Big Bubble DALK, although this premise was not held since adverse events occurred in both groups alike. A learning curve effect might have interfered this finding. The combination of a low trial inclusion rate and serious doubts on study safety prompted the termination of this trial in June 2015. We regard it our ethical obligation towards the participants to report these trial results nonetheless (Edwards et al., 1997).

Both treatments arms were confronted with a remarkably high rate of adverse events (AE), and these can be viewed from different perspectives. From a trial perspective, conversion to a perforating surgery was the most relevant AE. From an ethical/juridical perspective the AEs that require a re-operation, i.e., the detached Descemet membranes, could be considered the most severe. From a patient perspective, however, the AEs that negatively influence optimal visual acuity can be regarded the most burdensome, in other words, the Descemet folds that impair visual acuity on the long term. Apart from the intrinsic difficulties and long learning curve associated with lamellar surgery (Kasbekar et al., 2014), the degree of keratoconus in this study was very severe, with an average $K_{max}$ of 76.6D and an average pachymetry of 322µm. If these two mean values are considered a compound index of the staging of keratoconus severity, interesting comparisons can be made with other surgical studies (Anwar & Teichmann, 2002; Busin et al., 2012; Ghanem et al., 2015), should the baseline characteristics be adequately reported. The increased availability and the clinical experience with scleral contact lenses in the Netherlands can be considered a contributing factor for this difference: With adequately fitted scleral lenses virtually all clear keratoconus corneas can achieve a good visual acuity (Visser et al., 2016).

Another consideration is that the treatment protocol did not formally exclude scarred corneas. In clinical practice, however, lamellar surgery in these cases pertains an even higher risk of Descemet perforation/rupture, and lamellar surgery after a sustained hydrops is rarely successfully completed (Wisse et al., 2014). During the course of the trial, cases with a sustained corneal hydrops were not considered suitable for trial participation, mainly because performing a successful Descemet baring DALK becomes increasingly technically demanding. Secondly, these cases were considered unsuitable because the scarred residual stroma in a PET procedure might preclude optimal visual recovery. Apart from abovementioned alterations, the surgical and clinical protocol remained virtually unchanged. This could be considered a strength of this study, in the light of the difficult equilibrium between trial obligations and surgical innovation. Researchers have debated that the timeframe of a well-conducted trial spans many years (Beks et al., 2017); years in which the investigated technique can be adjusted and improved. What, then, is the value of a trial if it provides evidence based medicine for yesterday's procedures? The latter is of particular relevance in corneal surgery. Busin himself published an improved technique for keratoconus surgery which renders the previously reported PET technique obsolete (Busin et al., 2016).

In conclusion, a significant increase of uncorrected and corrected visual acuity was recorded for the group as a whole 1 year af-

ter corneal transplantation surgery for keratoconus. The added value of the PET over the DALK technique in terms of surgical safety cannot be deducted from these data, nor could we assess differences in the secondary outcomes (e.g., visual acuity, endothelial cell loss). However, in either treatment arm the incidence of intra-operative adverse events was higher than expected.

## Disclosure

None of the authors have any conflict of interest to disclose.

## Funding

This research was funded by an unrestricted grant from the Dr. F.P. Fisscher Stichting Utrecht, The Netherlands, facilitated by Stichting Vrienden van het UMC Utrecht.

## References

Amayem, A. F., & Anwar, M. (2000). Fluid lamellar keratoplasty in keratoconus. *Ophthalmology*, *107*(1), 76-79. http://doi.org/10.1016/S0161-6420(99)00002-0

Ang, L., Boruchoff, S., & Azar, D. T. (2009). Penetrating keratoplasty. In D. M. Albert, J. W. Miller, D. T. Azar, & B. A. Blodi (Eds.), *Albert & Jakobiec's Principles and practice of ophthalmology* (3rd ed., pp. 813-827). Elsevier.

Anwar, M., & Teichmann, K. D. (2002). Big-bubble technique to bare Descemet's membrane in anterior lamellar keratoplasty. *Journal of Cataract & Refractive Surgery, 28*(3), 398-403. http://doi.org/10.1016/S0886-3350(01)01181-6

Beks, R. B., Houwert, R. M., & Groenwold, R. H. H. (2017). Meerwaarde van observationeel onderzoek in chirurgie. *Nederlands Tijdschrift voor Geneeskunde, 161*, Article D1493.

Buratto, L., Ferrari, M., & Rama, P. (1992). Excimer laser intrastromal keratomileusis. *American Journal of Ophthalmology, 113*(3), 291-295. http://doi.org/10.1016/s0002-9394(14)71581-8

Busin, M., Scorcia, V., Leon, P., & Nahum, Y. (2016). Outcomes of air injection within 2 mm inside a deep trephination for deep anterior lamellar keratoplasty in eyes with keratoconus. *American Journal of Ophthalmology, 164*, 6-13. http://doi.org/10.1016/j.ajo.2015.12.033

Busin, M., Scorcia, V., Zambianchi, L., & Ponzin, D. (2012). Outcomes from a modified microkeratome-assisted lamellar keratoplasty for keratoconus. *Archives of Ophthalmology*, *130*(6), 776-782. http://doi.org/10.1001/archophthalmol.2011.1546

Cheng, Y. Y. Y., Visser, N., Schouten, J. S., Wijdh, R.-J., Pels, E., van Cleynenbreugel, H., Eggink, C. A., Zaal, M. J. W., Rijneveld, W. J., & Nuijts, R. M. M. A. (2011). Endothelial cell loss and visual outcome of deep anterior lamellar keratoplasty versus penetrating keratoplasty: A randomized multicenter clinical trial. *Ophthalmology, 118*(2), 302-309. http://doi.org/10.1016/j.ophtha.2010.06.005

Chow, S-C., Wang, H., & Shao, J. (Eds.). (2003). *Sample Size Calculations in Clinical Research*. Chapman & Hall/CRC.

D'Agostino, R. B., Chase, W., Belanger, A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial populations. *The American Statistician, 42*(3), 198-202. http://doi.org/10.1080/00031305.1988.10475563

Edwards, S. J. L., Lilford, R. J., Braumholtz, D., & Jackson, J. (1997). Why "underpowered" trials are not necessarily unethical. *The Lancet, 350*(9080), 804-807. http://doi.org/10.1016/S0140-6736(97)02290-3

European Eye Bank Association. (2008). Agreement on minimal standards (AMS). https:/ / www.eeba.eu/ files/ pdf / EEBA%20Minimum%20Medical%20Standards%20Revision%205%20Final.pdf

Folks, G. (2005). Diagnoses and management of corneal allograft rejection. In J. Krahmer, M. Mannis, & E. Holland (Eds.), *CORNEA* (2nd ed., pp. 1284-1314). Mosby.

Fontana, L., Parente, G., & Tassinari, G. (2007). Clinical outcomes after deep anterior lamellar keratoplasty using the big-bubble technique in patients with keratoconus. *American Journal of Ophthalmology, 143*(1), 117-124. http://doi.org/10.1016/j.ajo.2006.09.025

Ghanem, R. C., Bogoni, A., & Ghanem, V. C. (2015). Pachymetry-guided intrastromal air injection ("pachy-bubble") for deep anterior lamellar keratoplasty: Results of the first 110 cases. *Cornea*, *34*(6), 625-631. http://doi.org/10.1097/ICO.0000000000000413

Godefrooij, D. A., Gans, R., Imhof, S. M., &

Wisse, R. P. L. (2016). Nationwide reduction in the number of corneal transplantations for keratoconus following the implementation of crosslinking. *Acta Ophthalmologica, 94*(7), 675-678. http://doi.org/10.1111/aos.13095

Kasbekar, S. A., Jones, M. N. A., Ahmad, S., Larkin, D. F. P., & Kaye, S. B. (2013). Corneal transplant surgery for keratoconus and the effect of surgeon experience on deep anterior lamellar keratoplasty outcomes. *American Journal of Ophthalmology, 158*(6), 1239-1246. http://doi.org/10.1016/j.ajo.2014.08.029

Kennedy, R. H., Bourne, W. M., Dyer, J. A. (1986). A 48-year clinical and epidemiologic study of keratoconus. *American Journal of Ophtalmology, 101*(3), 267-273. http://doi.org/10.1016/0002-9394(86)90817-2

Melles, G. R., Remeijer, L., Geerards, A. J., Beekhuis, W., & Houdijn, M. D. (2000). A quick surgical technique for deep, anterior lamellar keratoplasty using visco-dissection. *Cornea*, *19*(4), 427-432. http:/ / doi.org/ 10.1097 /00003226-200007000-00004

Rabinowitz, Y. S., & Rasheed, K. (1999). KISA% index: A quantitative videokeratography algorithm embodying minimal topographic criteria for diagnosing keratoconus. *Journal of Cataract & Refractive Surgery, 25*(10), 1327-1335. http:/ / doi.org/ 10.1016/ S0886-3350(99)00195-9

Söğütlü Sari, E., Kubaloğlu, A., Ünal, M., Piñero Llorens, D., Koytak, A., Ofluoglu, A. N., & Özertürk, Y. (2012). Penetrating keratoplasty versus deep anterior lamellar keratoplasty: Comparison of optical and visual quality outcomes. *British Journal of Ophthalmology, 96*(8), 1063-1067. http://doi.org/10.1136/bjophthalmol-2011-301349

Tang, M., Shekhar, R., Miranda, D., & Huang, D. (2005). Characteristics of keratoconus and pellucid marginal degeneration in mean curvature maps. *American Journal of Ophthalmology, 140*(6), 993-1001. http://doi.org/10.1016/j.ajo.2005.06.026

Tolstoy, L. (1997). *Anna Karenina* (L. Maude & A. Maude, Trans.). Wordsworth Editions Limited. (Original work published in 1877)

Visser, E. S., Wisse, R. P. L., Soeters, N., Imhof, S. M., & van der Lelij, A. (2016). Objective and subjective evaluation of the performance of medical contact lenses fitted using a contact lens selection algorithm. *Contact Lens & Anterior Eye, 39*(4), 298-306. http://doi.org/10.1016/j.clae.2016.02.006

Wisse, R. P. L., van den Hoven, C. M. L., & van der Lelij, A. (2014). Does lamellar surgery for keratoconus experience the popularity it deserves? *Acta Ophthalmologica, 92*(5), 473-477. http://doi.org/10.1111/aos.12281

# Reflection on the PENTACON Trial: Lessons learned from an unpublished study

Robert P. L. Wisse[1]

As a clinician-scientist, the journey through the PENTACON trial, a multicenter randomized clinical study comparing two surgical techniques for corneal transplantation in keratoconus patients, was both enlightening and challenging. The study, which served as the centerpiece of my PhD research in 2015, ticked all the regulatory boxes and was ethically approved but ultimately fell short due to being gravely underpowered. This reflection aims to dissect the experiences, challenges, and lessons learned from this endeavor. Most importantly, while courses on clinical trial design such as Good Clinical Practice (GCP) certifications can teach important principles, the core value of any study remains somewhat elusive. How can one find the thin golden line between scientific value, innovation, patient recruitment, and overall trial feasibility? The outcomes of a study can never be completely sure; if we knew them ahead of time, it wouldn't be called research. However, in this particular example, there were tell-tale signs of failure early on. Therefore, I believe there are lessons for fellow researchers to be found.

**Keywords** *trial and error, clinical trial design, scientific equipoise, adversity, underpowered study*

[1] University Medical Center, Utrecht, the Netherlands

**Part of Special Issue**
Scientific Failure and Uncertainty in the Health Domain

**Received**
June 25, 2024
**Accepted**
April 27, 2025
**Published**
May 25, 2025

**Correspondence**
University Medical Center Utrecht
Rplwisse@gmail.com



## Personal and professional growth

The failure to publish the outcomes of the biggest endeavor of my PhD was a significant setback, but it also catalyzed my growth in unexpected ways. We conceptualized this clinical trial in 2010, based on the innovative and experimental work of Professor Busin in Forlí, Italia. His team developed many novel surgical approaches in corneal surgery, of which one innovation was particularly impactful: the microkeratome assisted anterior lamellar keratoplasty (Busin et al., 2012). Prof. Busin was awarded the "best of show" video session covering this technique by the American Academy of Ophthalmology annual conference in 2005, suggesting widespread appreciation of his approach in the field. Even today, he achieves scientific acclaim for his innovations in corneal surgery, as evidenced by top-tier publications considering the same type of procedures in the same population (see Bovone et al., 2024).

Our team in Utrecht was nimble, and this clinical trial was my first prospective project. My supervisor was a corneal surgeon and immunologist, with an established track record in immunological research, uveitis, and onchocerciasis. She performed these surgeries and collaborated with the Busin group. With great enthusiasm we embarked on this project, secured funding, and designed the clinical trial, dubbed the PENTACON. The public video library aiding surgeons in how to perform the surgery is still online (Utrecht Trial Videos, 2012).

### Companion Article

Robert P. L. Wisse (2025)

*Partial Endothelial Trepanation versus Deep Anterior Lamellar Keratoplasty in keratoconus patients: Results of the PENTACON trial*

DOI: 10.36850/0550-4e9c

The PENTACON report is clear; the trial never reached the needed participants to produce reliable scientific conclusions. After much postponement, we decided to cancel the trial due to insufficient recruitment. The adage "what doesn't kill you makes you stronger" resonated deeply with me during this period. The disappointment spurred me to explore other research avenues to meet the PhD requirements, leading to novel collaborations in corneal immunology and epidemiology. These new directions not only broadened my expertise but also reinforced my resilience and adaptability as a researcher. Yet sometimes I cannot resist the what-if question: What if we had been successful? Would my career be substantially different? Now, 10 years later, I'm not sure, but I think my other work compensated at least.

The experience of navigating the aftermath of the PENTACON trial taught me to value resilience and adaptability. This pivot was not just about salvaging a career path but about embracing the broader scope of scientific inquiry. I ventured into corneal immunology and epidemiology, fields that were not initially on my radar but have been an integral to my professional identity. This adaptability is a testament to the dynamic nature of scientific research, where setbacks can often lead to unanticipated breakthroughs and growth.

## The Intricacies of study design

One of the crucial lessons learned from the PENTACON trial was the essential nature of study design in clinical research. The trial's underpowered status highlighted the importance of seasoned advice from experienced mentors with a proven track record of successfully completing clinical trials. My supervisor at the time lacked this specific expertise, and as a junior researcher, I was not fully equipped to recognize this gap. When I got sound advice from a respected senior researcher, I didn't listen. He politely refused to let his center participate in this trial for the exact reason the study eventually failed: a too complex procedure in a niche population. Keratoconus patients are not particularly rare, as we've identified ourselves in an epidemiological study (Godefrooij et al., 2017). Suitable candidates needed to be advanced cases, otherwise a corneal transplant is too costly and high-risk, yet end-stage disease with scars and post-corneal hydrops was an exclusion criterion. The trouble of patient recruitment was apparent at the start, for those doctors willing to admit and see patients. Rather than consider their concerns, I listened to our external, seasoned surgeon based in Italy whose technique we planned to test. He was not involved enough in the study to provide the detailed feedback we needed. This experience underscored the necessity of having a robust support system and guidance from seasoned researchers who can anticipate and navigate the potential pitfalls of clinical trials.

This experience taught me that effective study design is not just about theoretical knowledge but also about practical wisdom. The guidance from mentors is invaluable. Their insights can often spell the difference between a study's success and failure. This trial was a stark reminder of the importance of listening to experienced voices and valuing their input, even when it challenges our preconceptions or plans.

## Challenges in multicenter studies

Based on the relative rarity of suitable candidates for trial enrollment, a multicenter study was inevitable. From a scientific perspective, this was also the preferred route, since multicenter study outcomes are often more generalizable and less prone to reflect local circumstances not controlled for during the study.

The multicenter nature of the PENTACON trial presented unique challenges, particularly with patient recruitment and center participation. Our largest anticipated center in Rotterdam failed to include a single patient despite extensive efforts to clear all ethical checks and barriers and several site visits. This experience taught me that not all centers will deliver as expected in a multicenter study. My advice: hedge for that event. Recruit more centers than are needed, since trial inclusion can run slow for a myriad of reasons. Effective communication, commitment, and follow-through from all participating centers are paramount for the success of multicenter trials. We painfully learned that (future) multicenter studies should carefully evaluate and select participating centers based on their demonstrated capability and commitment to deliver. In retrospect, the multicenter ap-

proach, while intended to enhance the trial's robustness, became one of its downfalls. The variability in center participation highlighted the critical need for stringent pre-trial assessments of each site's readiness and commitment. Several sites just included one or two patients, which further reduced the scientific value of the study. In summary, true multicenter studies involve more than ethical clearances and polite site visits; they require a deep dive into each center's patient population, logistical capabilities, and the willingness of local investigators to engage fully with the trial's demands. Investigators' personal networks are a huge asset in this aspect.

## Importance of feasibility

The ability to quickly include patients is a critical factor in the success of clinical trials. The PENTACON trial suffered from a slow inclusion rate, partly due to the rarity of eligible patients and the complexity of the surgical technique. This highlights the necessity of ensuring a steady and sufficient patient pool within the clinic or network before embarking on a trial. If a disease is rare, be aware. For early-stage career researchers, leveraging their network to gain more volume in inclusions sounds sensible, but is your network ready to deliver? Unless you have clear recruitment pathways, it may be more sensible to investigate an interesting disease using a different approach, capturing more data and delivering better conclusions.

The slow patient inclusion rate was a significant hurdle. In planning future trials, it is imperative to conduct thorough feasibility studies that accurately estimate the patient pool and the rate of inclusion. This involves not just statistical projections but also practical considerations about patient availability and the operational capacity of each site. A clear understanding of these factors can prevent the kind of slow inclusion that plagued the PENTACON trial and ultimately contributed to its downfall.

## Rigidity in protocols and its impact

Another significant lesson from the PENTACON trial was the impact of protocol rigidity over extended periods. The trial spanned several years, during which we wanted to update the surgical technique. Surgeons in this field are creative and always tinker; my supervisor was no exception. However, altering the protocol would have jeopardized the internal consistency of the study. This rigidity prevented iterative improvements and adaptations to the surgical procedure, ultimately contributing to the decision to terminate the trial. The relatively high rate of complications in both treatment arms in the PENTACON trial showed that the premise of a safe alternative was not met. As clinicians, we started to reject study participation for our patients, believing the technique under investigation to be inferior. Ergo, we lost our scientific equipoise: the genuine belief that both treatment options have equal value. This emphasized the need for a limited window of inclusions and adaptive trial designs that can accommodate advancements and refinements without compromising the study's integrity. Pallmann and colleagues (2018) advocate for adaptive trial methodologies that allow real-time adjustments, such as modifying sample sizes, adjusting randomization ratios, or incorporating interim analyses, thus enhancing the study's responsiveness without compromising validity. These factors can be anticipated in discussions with ethical boards prior and warrant input from experienced methodologists.

The tension between maintaining protocol fidelity and allowing for methodological advancements is a common challenge in clinical trials (Lawton et al., 2011). The PENTACON trial's rigidity prevented us from incorporating improvements that could have potentially salvaged the study. Future trials should consider more flexible designs that allow for adaptations in response to new developments, provided these changes are methodologically sound and ethically justified.

## Ethical obligation and reporting negative results

Despite the trial's termination and underpowered status, it was our ethical obligation to report the results. The manuscript became a chapter in my PhD (Wisse, 2015), and while it is not peer-reviewed, we can consider it publicly available (*grey literature*). Naturally, PhD dissertations are not indexed by search engines and scientific catalogs, and their outcomes are diffi-

cult to find. The PENTACON trial demonstrated that the PET technique was not as safe as anticipated, and both techniques showed a high rate of intra-operative complications. Although the article lacks detailed examples, our findings did highlight several conditions, such as the relatively high complication rates in both treatment arms, and the steep learning curve associated with the PET technique. The latter was not reported before in literature. Reporting these findings was crucial to inform the medical community and contribute to collective knowledge, preventing future researchers from encountering similar pitfalls. Especially at that time, the caveats and complications of these surgical techniques were gravely underreported in the scientific literature. No room for failure, apparently. This bias against failure has luckily changed, with rigorous and large studies being published that include the complications of complex corneal surgery (e.g., Feizi et al., 2023).

The ethical imperative to report all findings, regardless of their nature, cannot be overstated. Negative or inconclusive results are as crucial as positive ones in building a comprehensive understanding of medical interventions. Reluctance to publish negative findings perpetuates a skewed view of clinical efficacy and safety, leading to potential repetition of avoidable mistakes. Reporting bias is a known problem, jeopardizing the quality of our overall scientific endeavors and the value of funding research (Gill, 2012). By attempting to publish the results of the PENTACON trial, we aimed to fill a critical gap in the literature and foster a more transparent and informative scientific discourse. This was unfortunately not recognized by the peer-review process.

## Reflecting on broader impacts and future directions

The PENTACON trial, though unsuccessful in achieving its primary objectives, was a profound learning experience that shaped my career and research philosophy. It underscored the importance of robust study designs, realistic feasibility assessments, effective multicenter collaboration, protocol flexibility, and ethical reporting. These lessons have been invaluable in guiding my subsequent research endeavors and have contributed to my growth as a resilient and adaptive clinician-scientist.

Looking forward, the experiences from the PENTACON trial have prompted me to advocate for several key changes in how clinical trials are conducted and reported. Firstly, there is a need for more robust mentorship and collaboration frameworks that connect early-career researchers with seasoned investigators. Such frameworks can provide the necessary guidance and support to navigate the complexities of clinical research effectively.

Secondly, trial designs must incorporate flexibility to adapt to new developments and findings. Adaptive trial designs, which allow for modifications based on interim results, can enhance the relevance and applicability of clinical studies without compromising their integrity. This approach requires careful planning and ethical considerations but can significantly improve the efficiency and impact of clinical research.

Lastly, the scientific community must continue to emphasize the importance of publishing negative results. Journals and funding bodies should encourage the dissemination of all trial outcomes, fostering a more balanced and comprehensive understanding of medical interventions. This shift can help prevent the recurrence of past mistakes and guide future research more effectively.

## Conclusion

Through this reflection, I hope to share these insights with the broader medical community, fostering a culture of continuous learning and improvement in clinical research. The PENTACON trial, despite its shortcomings, has been a cornerstone in my development as a researcher. One learns the most from one's mistakes. The failed trial highlighted the intricacies of clinical trial design, the importance of experienced mentorship, the challenges of multicenter studies, and the ethical duty to report all findings. The lessons of this study have shaped my approach to research and will undoubtedly influence my future endeavors. Moreover, they serve as a reminder of the dynamic and often unpredictable nature of scientific inquiry, where every setback is an opportunity for growth and learning. I hope

this reflection can prevent fellow researchers from making these same mistakes.

## References

Bovone, C., De Rosa, L., Pellegrini, M., Ruzza, A., Ferrari, S., Camposampiero, D., Ponzin, D., Zauli, G., Yu, A. C., & Busin, M. (2024). Deep anterior lamellar keratoplasty using dehydrated versus standard organ culture-stored donor corneas: Prospective randomized trial. *Ophthalmology, 131*(6), 674-681. https://doi.org/10.1016/j.ophtha.2023.12.027

Busin, M., Scorcia, V., Zambianchi, L., & Ponzin, D. (2012). Outcomes from a modified microkeratome-assisted lamellar keratoplasty for keratoconus. *Archives of Ophthalmology, 130*(6), 776-782. https://doi.org/10.1001/archophthalmol.2011.1546

Feizi, S., Javadi, M. A., Karimian, F., Bayat, K., Bineshfar, N., & Esfandiari, H. (2023). Penetrating keratoplasty versus deep anterior lamellar keratoplasty for advanced stage of keratoconus. *American Journal of Ophthalmology, 248,* 107-115. https://doi.org/10.1016/j.ajo.2022.11.019

Gill, C. J. (2012). How often do US-based human subjects research studies register on time, and how often do they post their results? A statistical analysis of the Clinicaltrials.gov database. *BMJ Open, 2*(4), Article e001186. https://doi.org/10.1136/bmjopen-2012-001186

Godefrooij, D. A., de Wit, G. A., Uiterwaal, C. S., Imhof, S. M., & Wisse, R. P. (2017). Age-specific incidence and prevalence of keratoconus: A nationwide registration study. *American Journal of Ophthalmology, 175*, 169-172. https://doi.org/10.1016/j.ajo.2016.12.015

Lawton, J., Jenkins, N., Darbyshire, J. L., Holman, R. R., Farmer, A. J., & Hallowell, N. (2011). Challenges of maintaining research protocol fidelity in a clinical care setting: A qualitative study of the experiences and views of patients and staff participating in a randomized controlled trial. *Trials, 12*, Article 108. https://doi.org/10.1186/1745-6215-12-108

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Odoni, L., Sydes, M. R., Villar, S. S., Wason, J. M. S., Weir, C. J., Wheeler, G. M., Yap, C., & Jaki, T. (2018). Adaptive designs in clinical trials: Why use them, and how to run and report them. *BMC Medicine, 16*, Article 16. https://doi.org/10.1186/s12916-018-1017-7

Utrecht Trial Videos. (2012, May 3). *MALK for Keratoconus - The partial endothelial trepanation.* Vimeo. https://vimeo.com/37978511

Wisse, R. P. L. (2015). *Keratoconus : Inflammatory associations and treatment characteristics.* [Doctoral Dissertation, Utrecht University]. Utrecht University Repository. https://dspace.library.uu.nl/handle/1874/325112