

From Observation to a Justified Claim: Synthesis of Scientific Inference and Experimental Reasoning

Juan Segurola

3Dresyns by Resyner Technologies S.L.

© 2026 Juan Segurola. All Rights Reserved.

Correspondence: info@3Dresyns.com · www.3Dresyns.com

INF-013 · Scientific Inference and Experimental Reasoning · Version v0.2

Abstract

Scientific inference is not a single calculation performed after data collection. It is a controlled chain of commitments linking a stated claim to observations, assumptions, alternative explanations, tests, uncertainty and the target to which the conclusion is meant to apply. This synthesis integrates the ownership developed across INF-001 through INF-012 into one auditable reasoning sequence. The sequence begins by defining the claim and target, then asks how the data were generated and selected, what causal or mechanistic structure is assumed, which rival explanations remain compatible with the observations, how strongly the observed result discriminates among them, what prior knowledge legitimately constrains interpretation, which tests would expose error, how structural and parametric uncertainty are separated, whether conclusions survive defensible analytical variation, whether convergent evidence is genuinely independent in its failure modes, and whether the final conclusion can be transported beyond the conditions studied. The endpoint is not certainty. It is a claim whose scope, evidential basis, unresolved alternatives and residual uncertainty are explicit enough to be inspected and challenged.

Keywords: scientific inference; experimental reasoning; causal inference; evidence; identifiability; severe testing; uncertainty; robustness; triangulation; generalisation; justified claim

1. Inference is a chain, not an output

A numerical result can be correct while the scientific conclusion attached to it is unjustified. A precise estimate may answer the wrong question; an association may not identify a causal effect [1,2]; a fitted model may not be unique; a stable result may depend on unreported analytical choices; and an internally valid result may not apply to the intended target. INF therefore treats inference as a chain in which each link has its own failure conditions. The canonical order is not bureaucratic. It prevents later statistical or modelling sophistication from repairing an earlier ambiguity about what was actually claimed.

2. Start with the claim and the target

INF-001 owns the first discipline: descriptive, causal, mechanistic and predictive claims are different objects and require different evidence. Prediction can succeed without identifying a causal mechanism, and causal explanation cannot be inferred from predictive accuracy alone [4]. A defensible workflow therefore states the claim before selecting the evidence used to support it. It also states the target: the population, system, material state, intervention, boundary conditions, scale and time horizon to which the conclusion is intended to apply. Without a target, neither internal relevance nor later generalisation can be assessed.

3. Reconstruct how the observed data reached the analysis

Before interpreting an association, INF-002 and INF-003 require a structural account of data generation and selection. Confounders, mediators and colliders occupy different causal roles; conditioning on them can remove bias, block part of the effect of interest or create bias depending on the structure and estimand [1,2,5]. Selection and missingness add a second filter: the analysed observations may be a non-random subset of the states that could have been observed. The relevant question is not merely whether data are missing, but whether entry into the analysed set depends on variables connected to exposure, outcome or their common causes [5,6]. Measurement error can further distort effect estimates within this structure rather than simply adding symmetric noise [7].

4. Make rival explanations explicit

INF-004 separates observing a pattern from identifying the process that generated it. When several models, mechanisms or causal structures predict the same observations, the evidence is underdetermining. Structural identifiability asks whether the stated model and design determine the quantity or structure of interest uniquely from the observable distribution, even with unlimited data. Practical identifiability asks whether the finite data actually available constrain it usefully [8]. Auxiliary assumptions belong inside this comparison rather than outside it: calibration, boundary conditions, nuisance processes and model forms can rescue or defeat an apparent explanation. A justified claim therefore records what has been excluded, what remains viable and what evidence would discriminate among the remaining explanations.

5. Evidence strength is comparative

INF-005 and INF-006 distinguish the existence of a result from the amount by which it should alter a belief or preference among explanations. A p value is not the probability that a hypothesis is true, nor is it a direct measure of the size or practical importance of an effect [9]. Likelihood ratios and Bayes factors make the comparative character of evidence

explicit by asking how well observations are predicted under competing accounts [10,11]. Prior knowledge can legitimately constrain inference when it is explicit, relevant and open to challenge; it becomes a hidden source of bias when assumptions enter informally after the result is known. Base rates matter because the same evidential increment can lead to different posterior conclusions when starting plausibilities differ [11,12].

6. Prefer tests that could have exposed the explanation as wrong

INF-007 turns rival explanations into experimental leverage. A useful experiment is not merely one that produces another measurement; it produces outcomes that differ among the explanations being compared. Strong inference and severe testing share this discipline: design the test around errors or alternatives that the procedure has a real chance to reveal [3,13]. Failure to reject an explanation after an insensitive test is weak evidence. Conversely, an intervention that changes the predicted discriminator and the downstream outcome can strengthen causal attribution, provided the intervention does not simultaneously alter uncontrolled pathways. Operational falsification is therefore local and test-specific rather than a universal declaration that a theory has been permanently disproved.

7. Separate uncertainty in values from uncertainty in structure

INF-008 distinguishes parametric uncertainty within an accepted model from structural uncertainty about the model itself. Narrow confidence or posterior intervals cannot express uncertainty created by an omitted mechanism, wrong functional form, incorrect boundary condition or non-identifiable structure. Model calibration may also conflate parameter uncertainty with model discrepancy if the latter is not represented [14]. Sensitivity analysis is informative only when it varies the assumptions that matter for the conclusion; it cannot substitute for considering alternative structures that were never included [15].

8. Ask whether the conclusion survives defensible analytical variation

INF-009 and INF-010 address two different problems. Robustness analysis deliberately varies plausible assumptions, model forms or specifications and asks which conclusions survive. Specification-curve and multiverse approaches make this dependence visible when many defensible analyses exist [16,17]. Analytic multiplicity is a different risk, created when data exclusions, stopping rules, outcomes, covariates, transformations or models are selected after seeing results. Undisclosed flexibility can greatly inflate the chance of presenting a favourable result as if it were uniquely determined by the data [18]. Preregistration can constrain some forms of this flexibility [19], but it does not make a poor design informative or eliminate the need for sensitivity analysis.

9. Integrate evidence by independence of weaknesses, not by counting studies

INF-011 treats triangulation as structured convergence. Several measurements are most informative when their biases, assumptions and failure modes differ. Agreement among methods that share the same unrecognised bias can create false reassurance, while disagreement among methods can reveal which assumption needs examination. Triangulation therefore asks whether independent routes constrain the same claim from different directions, not how many publications report a similar answer [20,21]. Evidence integration remains claim-specific: measurements supporting prediction, mechanism and causality are not interchangeable merely because they concern the same system.

10. Generalisation is another inference

INF-012 closes the chain by separating validity in the studied conditions from applicability to a target. Generalisation and transportability require explicit assumptions about effect modification, support and which relationships remain stable between study and target [22,23]. Reweighting or modelling can bridge distributions only where the necessary variables are measured and the target lies within adequate support. Extrapolation into unrepresented states depends increasingly on structural assumptions. In physical and engineering systems the same logic applies to scale, geometry, process history, equipment state and environment: procedural replication alone does not establish that the same effect or mechanism governs the new target.

11. A minimum record for a justified scientific claim

- State the claim type and the target to which it applies.
- Describe how observations were generated, measured, selected and excluded.
- Make the causal or mechanistic structure needed by the claim explicit.
- List credible rival explanations and identify what remains non-identifiable.
- State what the observed result contributes relative to those alternatives and to relevant prior knowledge.
- Identify a discriminating or severe test and specify which error it is capable of revealing.
- Separate parametric uncertainty from structural uncertainty and model discrepancy.
- Test whether the conclusion survives defensible assumptions and analytical specifications.
- Integrate additional evidence by considering independence of biases and failure modes.
- Justify any transfer from studied conditions to the target and mark extrapolation explicitly.
- Report the strongest conclusion supported by the evidence, together with unresolved alternatives and residual uncertainty.

12. Stopping rule: justified does not mean final

A scientific conclusion is justified to the extent that it is stated no more strongly than the evidence discriminates it from the relevant alternatives within the declared scope and assumptions, and to the extent that the principal routes by which it

could be artefactual have been made explicit and tested by procedures capable of revealing them. Where the available methods lack that capability, the claim is narrowed accordingly rather than declared justified. This standard does not require certainty and does not imply permanence. New evidence can alter the causal structure, expose selection, reveal a previously unmodelled mechanism, change the evidential balance or show that the result does not transport. The correct stopping rule is therefore evidential rather than rhetorical: stop at the strongest claim the present evidence can carry, not at the most complete story that can be written.

13. Synthesis

The INF architecture converts experimental reasoning into an inspectable sequence. INF-001 defines the claim; INF-002 and INF-003 establish structural and selection conditions; INF-004 decides which rival explanations remain live; INF-005 and INF-006 discipline evidential updating; INF-007 demands tests capable of discrimination; INF-008 separates kinds of uncertainty; INF-009 and INF-010 expose dependence on analytical choices; INF-011 integrates complementary evidence; and INF-012 defines the conditions for transfer to a target. The series therefore ends where a scientific paper or technical decision should begin its conclusion: with a claim whose meaning, evidence, assumptions, alternatives, uncertainty and domain of validity are all visible. The value of the framework is not that it removes judgement, but that it makes the scientific grounds for that judgement traceable.

References

- [1] Pearl, J. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, 2009. DOI: 10.1017/CBO9780511803161.
- [2] Hernán, M.A.; Robins, J.M. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020. DOI: 10.1201/9781315374932.
- [3] Mayo, D.G. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, 2018. DOI: 10.1017/9781107286184.
- [4] Shmueli, G. To Explain or to Predict? *Statistical Science* 25(3), 289-310 (2010). DOI: 10.1214/10-STS330.
- [5] Hernán, M.A.; Hernández-Díaz, S.; Robins, J.M. A Structural Approach to Selection Bias. *Epidemiology* 15(5), 615-625 (2004). DOI: 10.1097/01.ede.0000135174.63482.43.
- [6] Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*. 3rd ed. Wiley, 2019. DOI: 10.1002/9781119482260.
- [7] Carroll, R.J.; Ruppert, D.; Stefanski, L.A.; Crainiceanu, C.M. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Chapman & Hall/CRC, 2006. DOI: 10.1201/9781420010138.
- [8] Raue, A.; Kreutz, C.; Maiwald, T.; Bachmann, J.; Schilling, M.; Klingmüller, U.; Timmer, J. Structural and Practical Identifiability Analysis of Partially Observed Dynamical Models by Exploiting the Profile Likelihood. *Bioinformatics* 25(15), 1923-1929 (2009). DOI: 10.1093/bioinformatics/btp358.
- [9] Greenland, S.; Senn, S.J.; Rothman, K.J.; Carlin, J.B.; Poole, C.; Goodman, S.N.; Altman, D.G. Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations. *European Journal of Epidemiology* 31, 337-350 (2016). DOI: 10.1007/s10654-016-0149-3.
- [10] Royall, R. *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall, 1997. DOI: 10.1201/9780203738665.
- [11] Kass, R.E.; Raftery, A.E. Bayes Factors. *Journal of the American Statistical Association* 90(430), 773-795 (1995). DOI: 10.1080/01621459.1995.10476572.
- [12] Greenland, S. Bayesian Perspectives for Epidemiological Research: I. Foundations and Basic Methods. *International Journal of Epidemiology* 35(3), 765-775 (2006). DOI: 10.1093/ije/dyi312.
- [13] Platt, J.R. Strong Inference. *Science* 146(3642), 347-353 (1964). DOI: 10.1126/science.146.3642.347.
- [14] Kennedy, M.C.; O'Hagan, A. Bayesian Calibration of Computer Models. *Journal of the Royal Statistical Society: Series B* 63(3), 425-464 (2001). DOI: 10.1111/1467-9868.00294.
- [15] Saltelli, A.; Ratto, M.; Andres, T.; Campolongo, F.; Cariboni, J.; Gatelli, D.; Saisana, M.; Tarantola, S. *Global Sensitivity Analysis: The Primer*. Wiley, 2008. DOI: 10.1002/9780470725184.
- [16] Simonsohn, U.; Simmons, J.P.; Nelson, L.D. Specification curve analysis. *Nature Human Behaviour* 4, 1208-1214 (2020). DOI: 10.1038/s41562-020-0912-z.
- [17] Steegen, S.; Tuerlinckx, F.; Gelman, A.; Vanpaemel, W. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* 11(5), 702-712 (2016). DOI: 10.1177/1745691616658637.
- [18] Simmons, J.P.; Nelson, L.D.; Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22(11), 1359-1366 (2011). DOI: 10.1177/0956797611417632.
- [19] Nosek, B.A.; Ebersole, C.R.; DeHaven, A.C.; Mellor, D.T. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115(11), 2600-2606 (2018). DOI: 10.1073/pnas.1708274114.
- [20] Lawlor, D.A.; Tilling, K.; Davey Smith, G. Triangulation in aetiological epidemiology. *International Journal of Epidemiology* 45(6), 1866-1886 (2016). DOI: 10.1093/ije/dyw314.
- [21] Munafò, M.R.; Davey Smith, G. Robust research needs many lines of evidence. *Nature* 553, 399-401 (2018). DOI: 10.1038/d41586-018-01023-3.

[22] Dahabreh, I.J.; Robertson, S.E.; Tchetgen Tchetgen, E.J.; Stuart, E.A.; Hernán, M.A. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* 75(2), 685-694 (2019). DOI: 10.1111/biom.13009.

[23] Dahabreh, I.J.; Robertson, S.E.; Steingrimsson, J.A.; Stuart, E.A.; Hernán, M.A. Extending inferences from a randomized trial to a new target population. *Statistics in Medicine* 39(14), 1999-2014 (2020). DOI: 10.1002/sim.8426.

Data and materials statement

No new experimental dataset was generated for this article. The article is a methodological synthesis based on the cited literature and on the frozen ownership architecture of INF-001 through INF-012.