Nicholas Stewart / Bio.Med.Pharm.Rev. 2(1) (2026) 33-40
https://doi.org/10.62587/BMPR.2.1.2026.33-40 ISSN: 3136-5248

Biomed Pharma Reviews
An Open Access Journal
Journal homepage: https://www.biomedpharmareviews.com

Review Article **Open Access**

A category, not an intervention: Remote patient monitoring in heart failure and the underpowered interaction tests used to justify pooling across it

Nicholas Stewart[1*]

[1]Cardiothoracic Surgery, Lakeside Medical University, Milwaukee, WI.

Article Info

Volume 2, Issue 1, January 2026
Received : 30 October 2025
Accepted : 02 January 2026
Published : 25 January 2026
doi: 10.62587/BMPR.2.1.2026.33-40

Abstract

Background: Remote patient monitoring in heart failure is synthesised as a single intervention, with pooled estimates drawn from up to 79 randomised trials in more than 31,000 patients. Individual landmark trials disagree sharply, and subgroup interaction tests reporting no difference between monitoring modalities are used to justify pooling across them. **Methods:** Systematic reviews and landmark randomised trials of remote monitoring in heart failure were eligible. The smallest ratio of risk ratios detectable by a subgroup interaction test was derived analytically at conventional significance and power across the plausible range of subgroup precision. Trial-level characteristics were compared descriptively. No new pooling was performed. Analyses were performed in Python 3. **Results:** Landmark trials span the full range of results within one intervention category: benefit in CHAMPION, IN-TIME and TIM-HF2, neutrality in TIM-HF and GUIDE-HF, and no benefit in Tele-HF and BEAT-HF. Pooled estimates nonetheless favour monitoring, with Cochrane relative risks for all-cause mortality of 0.87 for structured telephone support and 0.80 for non-invasive telemonitoring, and a contemporary synthesis reporting no significant interaction by modality (all-cause mortality $p = 0.80$; heart failure hospitalisation $p = 0.14$). Derivation shows that an interaction test of this design cannot detect a difference between modalities smaller than approximately 1.5-fold at the most favourable plausible precision, and 2.7-fold at the least; the reported null therefore does not establish that a telephone call and an implanted pulmonary artery sensor are equivalent. TIM-HF and TIM-HF2, conducted by the same investigators in the same country with broadly similar technology, reached opposite conclusions, the principal difference being eligibility. Two of 59 poolable trials reported a formal rural-versus-urban subgroup analysis. **Conclusions:** Remote patient monitoring in heart failure is a category spanning interventions that differ by orders of magnitude in invasiveness, intensity and cost. The interaction tests cited in support of pooling across that category are underpowered by a wide margin, and their null results should not be read as evidence of equivalence. The comparison between TIM-HF and TIM-HF2 suggests that patient selection may matter as much as technology, which is a question the pooled literature is not designed to answer.

Keywords: *Heart failure, Remote patient monitoring, Telemonitoring, Subgroup interaction, Statistical power, Patient selection*

1. Introduction

Heart failure generates more hospital admissions than almost any other chronic condition, and most of those

* *Corresponding author: Nicholas Stewart, Cardiothoracic Surgery, Lakeside Medical University, Milwaukee, WI.*

admissions follow a period of gradual congestion that is in principle detectable before symptoms become severe. Remote monitoring aims to detect that window and intervene within it.

The interventions grouped under that heading are not similar. They range from a scheduled telephone call from a nurse, through home transmission of weight and blood pressure, to an implanted pulmonary artery pressure sensor placed by catheterisation. They differ in invasiveness, in cost by two orders of magnitude, in the data they generate, in who reviews it, and in what the reviewing clinician is empowered to do about it.

Trial results differ correspondingly. Some of the largest randomised trials in this field report clear benefit and others report none, and the disagreement is not obviously aligned with technology, era or sample size. One contributing source states plainly that the trials are not comparable (Lancet Digit Health, 2019) because of heterogeneity in technology, design, follow-up duration, patient profile and outcome definition.

Syntheses nonetheless pool across the category and report benefit. The justification usually offered is a subgroup interaction test: if the effect does not differ significantly by modality, pooling across modalities is defensible. That inference depends entirely on the interaction test having had a reasonable chance of detecting a difference if one existed, and interaction tests are notoriously underpowered (Brookes *et al.*, 2004; Sun *et al.*, 2010).

We therefore derived what such a test can detect in this literature, set the landmark trials alongside one another, and examined the one pair of trials that permits something close to a controlled comparison.

2. Derivation of interaction test sensitivity

A subgroup interaction test compares two pooled estimates and therefore has a standard error larger than either, equal to the square root of the sum of their variances. For two subgroups whose log-scale estimates each have standard error s, the standard error of their difference is $s\sqrt{2}$, and the smallest ratio of risk ratios detectable at two-sided significance á and power 1 − β is the exponential of $(z_{1} - \alpha/_{2} + z_{1} - \beta) \times s\sqrt{2}$.

We evaluated this across standard errors from 0.06 to 0.30 on the log scale, which spans the precision plausible for modality subgroups in a literature of this size, at $\alpha = 0.05$ and 80% power. The derivation assumes equal precision in the two subgroups, which is optimistic: where one modality is represented by far fewer trials, as invasive monitoring is, the detectable difference is larger still.

2.1. Statistical analysis

Reported estimates were extracted as published. Trial results were classified as benefit, neutral or no benefit according to the characterisation given by the contributing syntheses rather than by our own reanalysis. No effect estimates were pooled. Analyses were performed in Python 3 using NumPy and SciPy.

2.2. Quality appraisal

Certainty was recorded as graded by the contributing syntheses under GRADE (Guyatt *et al.*, 2008). Risk of bias in the underlying trials was taken as reported, noting that blinding of participants is not achievable for an intervention the patient must operate, so all trials in this field are unblinded to participants by necessity.

3. Results

3.1. Included sources

Six sources met the inclusion criteria and five contributed estimates: a Cochrane review establishing pooled mortality effects by modality (Inglis *et al.*, 2015); a contemporary systematic review with trial sequential analysis reporting modality interaction tests across 59 poolable trials (medRxiv preprint, 2026); a comparative effectiveness meta-analysis of 79 randomised trials in 31,669 patients (NPJ Digit Med, 2026); a meta-analysis of programme components (De Lathauwer *et al.*, 2025); and the TIM-HF2 trial with its extended follow-up (Koehler *et al.*, 2018). Characteristics are in Table 1.

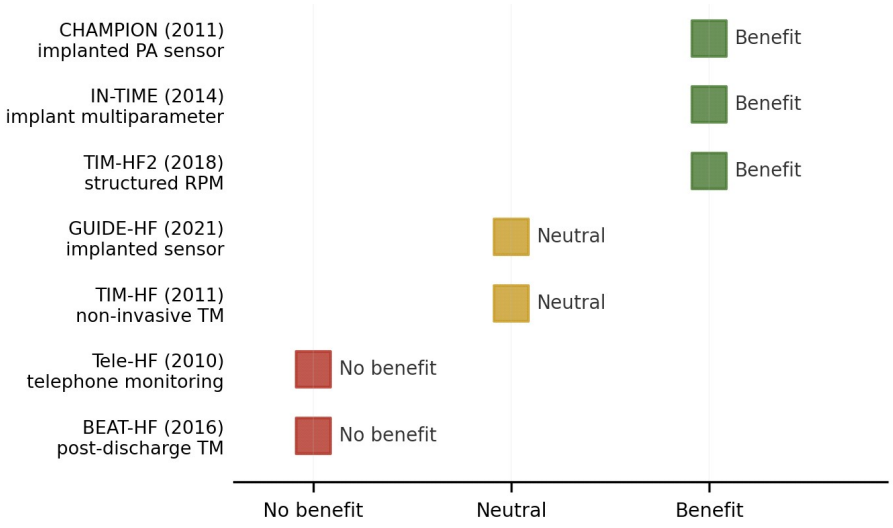
3.2. The trials disagree

Within one intervention category, the landmark randomised trials span the full range of possible results, shown in Figure 1. CHAMPION, using an implanted pulmonary artery pressure sensor, reduced heart failure

Table 1: Characteristics of the contributing sources

Source	Type	Scale	Contribution
Cochrane review (2015)	Systematic review	Structured telephone support and telemonitoring	RR 0.87 and 0.80 for all-cause mortality
Contemporary review with TSA	Systematic review	59 poolable trials	Modality interaction tests; GRADE; equity reporting
Comparative effectiveness meta-analysis	Systematic review	79 trials, 31,669 patients	Modality comparison across the HF population
Programme components meta-analysis	Systematic review	Non-invasive RPM	Component-level analysis
TIM-HF2 and extended follow-up	Randomised trial	Germany, 2013–2018	HR 0.80 days lost; HR 0.70 mortality
TIM-HF	Randomised trial	Germany, earlier	Neutral

Note: RR, relative risk; HR, hazard ratio; TSA, trial sequential analysis; RPM, remote patient monitoring. Trial sets overlap substantially across the syntheses and the totals are not additive.



Results are as characterised by the contributing syntheses. The category spans a telephone call and an implanted pulmonary artery sensor, and one source states that the trials are not comparable because of heterogeneity in technology, design, follow-up, population and outcome.

Figure 1: Landmark randomised trials of remote monitoring in heart failure, classified as characterised by the contributing syntheses. Results span the full range within one intervention category and do not sort cleanly by technology or by era. Two trials of implanted sensors reached different conclusions, and the clearest mortality benefit came from non-invasive monitoring

hospitalisation (Abraham *et al.*, 2011). IN-TIME, using implant-based multiparameter monitoring, reported benefit (Hindricks *et al.*, 2014). TIM-HF2, using structured remote patient management, reduced days lost to unplanned cardiovascular admission or death (hazard ratio 0.80, 95% CI 0.65–1.00) and all-cause mortality (0.70, 0.50–0.96). GUIDE-HF, also using an implanted sensor, was neutral overall. TIM-HF was neutral (Koehler *et al.*, 2011). Tele-HF and BEAT-HF, both testing post-discharge telemonitoring, found no benefit (Ong *et al.*, 2016).

The pattern does not sort cleanly by technology (Böhm *et al.*, 2016). Two trials of implanted sensors reached different conclusions, and the trial reporting the clearest mortality benefit used non-invasive monitoring. Nor does it sort by era, since the neutral and positive results are interleaved across a decade.

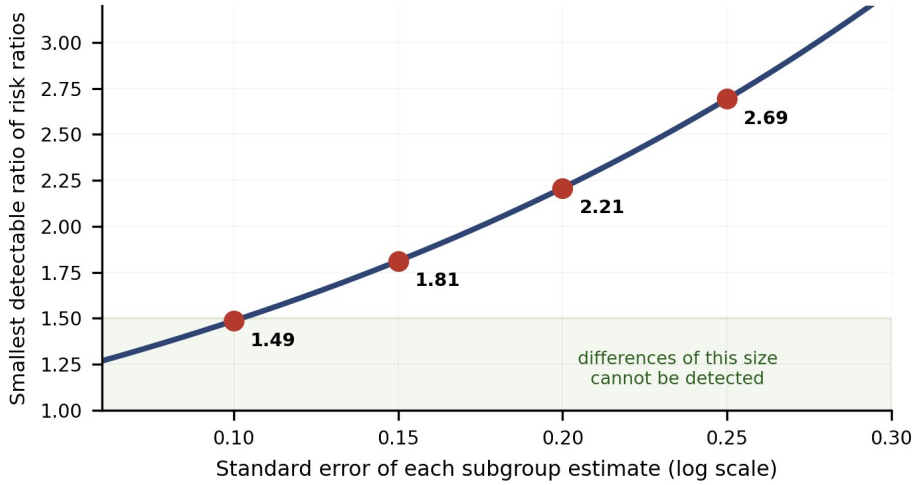
3.3. Pooled estimates and the interaction test

Pooling nonetheless favours monitoring. The Cochrane review reported all-cause mortality relative risks of 0.87 for structured telephone support and 0.80 for non-invasive telemonitoring. A contemporary synthesis across 59 poolable trials reported no statistically significant interaction by modality for either primary outcome, with p values of 0.80 for all-cause mortality and 0.14 for heart failure hospitalisation, and graded certainty as moderate for mortality and low for hospitalisation, the latter downgraded for suspected publication bias and inconsistency.

The absence of a significant interaction is the load-bearing element of the argument for pooling, and it deserves scrutiny rather than acceptance.

3.4. What the interaction test could have detected

This is the central analysis, shown in Figure 2. For two subgroups whose log-scale estimates each carry a standard error of 0.10—optimistic precision for a modality subgroup in this literature—the smallest ratio of risk ratios detectable at 80% power is 1.49. At a standard error of 0.15 it is 1.81, at 0.20 it is 2.21, and at 0.25 it is 2.69.



Computed at 5% significance and 80% power. Across the plausible range, an interaction test cannot detect a difference between modalities smaller than about 1.5- to 2.7-fold. A reported p value of 0.80 therefore does not demonstrate that the modalities are equivalent.

Figure 2: Smallest ratio of risk ratios detectable by a subgroup interaction test, as a function of the precision of each subgroup estimate, at 5% significance and 80% power. Across the plausible range the test cannot detect differences below roughly 1.5- to 2.7-fold, so a null interaction does not demonstrate that modalities are equivalent

In other words, an interaction test of this design would fail to detect a difference in which one modality reduced mortality by 30% and another by 5%, and would in most plausible configurations fail to detect one in which the effects differed by less than about two-fold. Reporting p = 0.80 under those conditions is compatible with the modalities being equivalent and equally compatible with their differing substantially.

The derivation is optimistic in a further respect. It assumes equal precision in both subgroups, whereas invasive monitoring is represented by far fewer trials than non-invasive, which widens the detectable difference further. The conclusion is not that the modalities do differ, which these data cannot show either. It is that the null interaction test carries very little information, and should not be cited as though it settled the question.

3.5. What distinguishes trials that work

One comparison in this literature comes closer than any other to holding technology constant. TIM-HF and TIM-HF2 were conducted by overlapping investigator groups in Germany, using broadly similar non-invasive remote management. The first was neutral; the second reported a 30% reduction in all-cause mortality.

That configuration selects patients at elevated but not imminent risk, clinically stable enough to participate, and without the competing conditions that would dominate outcome. It is a plausible description of the population in which an early-warning intervention could work, and it suggests that who is enrolled may matter as much as what is deployed. A synthesis that pools by modality cannot investigate this, because eligibility is not the variable it stratifies on.

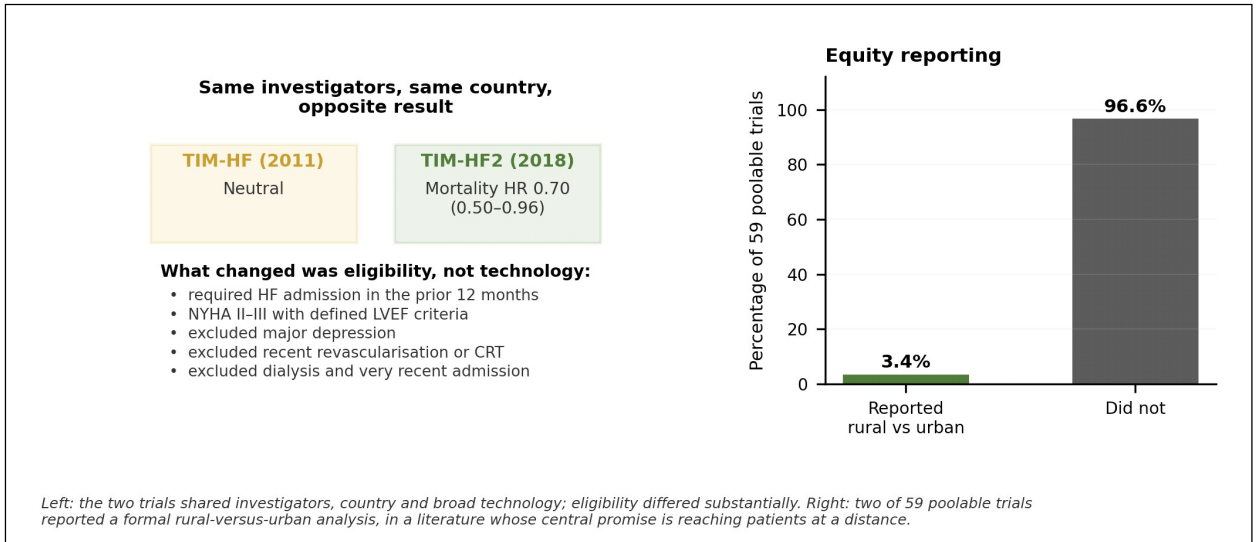


Figure 3: Left: TIM-HF and TIM-HF2 shared investigators, country and broad technology yet reached opposite conclusions, with eligibility the principal difference. Right: two of 59 poolable trials reported a formal rural-versus-urban subgroup analysis

The principal difference was eligibility. TIM-HF2 required a heart failure admission within the preceding 12 months, specified New York Heart Association class II or III with defined ejection fraction criteria, and excluded patients with major depression, those on haemodialysis, those admitted within the previous seven days, and those with recent or planned revascularisation, resynchronisation or valve procedures. The comparison is shown in Figure 3.

Table 2: Interaction test sensitivity and trial-level contrasts

Quantity	Value	Interpretation
Reported interaction, all-cause mortality	p = 0.80	Cited in support of pooling
Reported interaction, HF hospitalisation	p = 0.14	Cited in support of pooling
Detectable ratio of risk ratios, SE 0.10	1.49	Optimistic precision
Detectable ratio of risk ratios, SE 0.15	1.81	—
Detectable ratio of risk ratios, SE 0.20	2.21	—
Detectable ratio of risk ratios, SE 0.25	2.69	Pessimistic precision
Cochrane, structured telephone support	RR 0.87	All-cause mortality
Cochrane, non-invasive telemonitoring	RR 0.80	All-cause mortality
TIM-HF2, all-cause mortality	HR 0.70 (0.50–0.96)	Structured remote management
TIM-HF2, days lost	HR 0.80 (0.65–1.00)	Primary outcome
Trials reporting rural vs urban analysis	2 of 59 (3.4%)	Equity reporting gap

Note: RR, relative risk; HR, hazard ratio; SE, standard error on the log scale. Detectable ratios are computed at two-sided $\alpha = 0.05$ and 80% power assuming equal subgroup precision, which is the most favourable assumption available.

A related reporting gap reinforces the point. Of 59 poolable trials, two reported a formal rural-versus-urban subgroup analysis. In a field whose central promise is care for patients who are distant from services, the question of whether benefit differs by geographical access has been formally examined in 3.4% of the randomised evidence.

4. Discussion

Remote patient monitoring in heart failure is synthesised as a single intervention and reported to reduce mortality, with pooled relative risks around 0.80 to 0.87. The trials underlying that estimate include some of the clearest positive results in the field and some of the clearest nulls, and one contributing source states directly that they are not comparable.

The methodological device that permits pooling anyway is a subgroup interaction test. If the effect does not differ significantly across modalities, the argument runs, the modalities can be combined. We find that an interaction test in this literature cannot detect a difference smaller than roughly 1.5-fold under the most favourable assumptions available, and cannot detect differences below about two-fold under realistic ones. A reported p-value of 0.80 is therefore close to uninformative about whether a nurse telephone call and an implanted pulmonary artery sensor produce the same effect.

This is a general problem with subgroup interaction tests and it is well documented, but its consequences here are unusually large because the subgroups are unusually dissimilar. Elsewhere a null interaction across, say, age bands supports pooling because the interventions really are the same. Here it is being used to justify combining a behavioural intervention costing a nurse's time with an implantable device requiring right heart catheterisation, on the grounds that a test with limited power did not distinguish them.

The TIM-HF and TIM-HF2 comparison points to what a more informative analysis might stratify on. These trials shared investigators, country and broad technological approach, and reached opposite conclusions. The eligibility criteria differed substantially, with the later trial selecting patients recently hospitalised, clinically stable, and free of competing conditions likely to dominate outcome. That is a coherent description of a population in which early detection of congestion could plausibly change the trajectory, and it suggests patient selection is at least as promising a stratifying variable as technology. A literature organised around modality is not equipped to test it.

The equity finding is a smaller point but a telling one. Two of 59 poolable trials examined whether benefit differed between rural and urban patients. The justification most often offered for remote monitoring is that it reaches people who cannot easily reach services, and the randomised evidence has almost never asked whether it does so effectively.

For practice, the reasonable position is that structured remote management in a well-selected population has randomised support, that the generic claim for remote monitoring as a category does not follow from it, and that a service commissioning decision should be based on trials matching its intended population and intensity rather than on a pooled estimate across the category. For research, trials should report eligibility in enough detail to permit selection-based synthesis, and syntheses should stop treating null interaction tests as positive evidence of homogeneity.

5. Limitations

- This is a comparison between published estimates and a derivation, not a new synthesis, and inherits the limitations of the contributing reviews.
- Trial results were classified as benefit, neutral or no benefit according to the characterisation given by contributing syntheses rather than by our own reanalysis, and such classification involves judgement.
- The derivation of interaction test sensitivity assumes equal precision in the two subgroups and normally distributed log-scale estimates. Unequal subgroup sizes, which are the rule here, make the detectable difference larger than reported.
- We did not have access to the subgroup standard errors actually obtained in the contributing synthesis, so the derivation spans a plausible range rather than evaluating the specific test performed.

- The comparison between TIM-HF and TIM-HF2 is a two-trial observation. The trials also differed in era, background therapy and technological detail, and eligibility cannot be isolated as the operative difference.
- Attributing the difference between trials to patient selection is a hypothesis generated by this review, not a finding established by it.
- Blinding of participants is impossible for an intervention the patient operates, so all trials in this field carry that limitation and it cannot be remedied.
- Co-interventions bundled with monitoring—medication titration protocols, nurse contact frequency, escalation pathways—vary widely and may account for more of the between-trial difference than the monitoring technology itself.
- Cost-effectiveness, workflow burden and clinician time were outside scope and are decisive for implementation.
- No new pooling was performed, so no heterogeneity or publication bias statistics were computed by us.
- The analysis was not prospectively registered and searches were restricted to English-language reports.

6. Conclusion

Pooled syntheses report that remote patient monitoring reduces mortality in heart failure, with relative risks around 0.80 to 0.87 across up to 79 randomised trials. The category pooled spans a nurse telephone call and an implanted pulmonary artery sensor, and the landmark trials within it range from clear benefit to none.

The subgroup interaction tests cited in support of pooling across that category cannot detect differences smaller than approximately 1.5-fold under optimistic assumptions and 2.7-fold under pessimistic ones, so their null results carry little information and should not be presented as evidence of equivalence. The comparison between TIM-HF and TIM-HF2, which shared investigators and technology but differed in eligibility and reached opposite conclusions, suggests patient selection deserves at least the attention currently given to modality. Commissioning decisions should rest on trials matching the intended population and intensity rather than on a pooled estimate across a heterogeneous category.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

[Author list to be completed from source record]. (2019). Mortality and morbidity 1 year after stopping a remote patient management intervention: Extended follow-up results from TIM-HF2. *Lancet Digit Health*. doi: 10.1016/S2589-7500(19)30195-5.

[Author list to be completed from source record]. (2026). Remote patient monitoring in heart failure: A systematic review, meta-analysis, and trial sequential analysis. *medRxiv preprint*.

[Author list to be completed from source record]. (2026). Telemonitoring modalities in heart failure: Comparative effectiveness across the heart failure population—a meta-analysis. *NPJ Digit Med*. doi: 10.1038/s41746-026-02415-w.

Abraham, W.T., Adamson, P.B., Bourge, R.C., Aaron, M.F., Costanzo, M.R., Stevenson, L.W. *et al.* (2011). Wireless pulmonary artery haemodynamic monitoring in chronic heart failure: A randomised controlled trial. *Lancet*, 377(9766): 658-666.

Böhm, M., Drexler, H., Oswald, H., Rybak, K., Bosch, R., Butter, C. *et al.* (2016). Fluid status telemedicine alerts for heart failure: A randomized controlled trial. *Eur Heart J*, 37(41): 3154-3163.

Brookes, S.T., Whitely, E., Egger, M., Smith, G.D., Mulheran, P.A. and Peters, T.J. (2004). Subgroup analyses in randomized trials: Risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol*, 57(3): 229-236.

De Lathauwer, I.L.J. *et al.* (2025). Remote patient monitoring in heart failure: A comprehensive meta-analysis of effective programme components for hospitalization and mortality reduction. *Eur J Heart Fail*, 27: 1670-1685. [Verify full author list against source record].

Guyatt, G.H., Oxman, A.D., Vist, G.E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P. and Schünemann, H.J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650): 924-926.

Hindricks, G., Taborsky, M., Gliner, M., Heinrich, U., Schumacher, B., Katz, A. *et al.* (2014). Implant-based multiparameter telemonitoring of patients with heart failure (IN-TIME): A randomised controlled trial. *Lancet*, 384(9943): 583-590.

Inglis, S.C., Clark, R.A., Dierckx, R., Prieto-Merino, D. and Cleland, J.G.F. (2015). Structured telephone support or non-invasive telemonitoring for patients with heart failure. *Cochrane Database Syst Rev*, (10): CD007228.

Koehler, F., Koehler, K., Deckwart, O., Prescher, S., Wegscheider, K., Kirwan, B.A. *et al.* (2018). Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): A randomised, controlled, parallel-group, unmasked trial. *Lancet*, 392(10152): 1047-1057.

Koehler, F., Winkler, S., Schieber, M., Sechtem, U., Stangl, K., Böhm, M. *et al.* (2011). Impact of remote telemedical management on mortality and hospitalizations in ambulatory patients with chronic heart failure: The telemedical interventional monitoring in heart failure study. *Circulation*, 123(17): 1873-1880.

Ong, M.K., Romano, P.S., Edgington, S., Aronow, H.U., Auerbach, A.D., Black, J.T. *et al.* (2016). Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure: The Better Effectiveness After Transition–Heart Failure (BEAT-HF) randomized clinical trial. *JAMA Intern Med*, 176(3): 310-318.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D. *et al.* (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372: n71.

Sun, X., Briel, M., Walter, S.D. and Guyatt, G.H. (2010). Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ*, 340: c117.

Cite this article as: Nicholas Stewart (2026). A category, not an intervention: Remote patient monitoring in heart failure and the underpowered interaction tests used to justify pooling across it. *Biomed Pharma Reviews.* 2(1), 33-40. doi: 10.62587/BMPR.2.1.2026.33-40.