SCIENTIFIC REPORTS

OPEN

Received: 14 September 2017
Accepted: 9 April 2018
Published online: 16 May 2018

Long term extension of a randomised controlled trial of probiotics using electronic health records

Gareth Davies[1], Sue Jordan[2], Caroline J. Brooks[1], Daniel Thayer[1], Melanie Storey[2], Gareth Morgan[3,4], Stephen Allen[5], Iveta Garaiova[6], Sue Plummer[6] & Mike Gravenor[1]

Most randomised controlled trials (RCTs) are relatively short term and, due to costs and available resources, have limited opportunity to be re-visited or extended. There is no guarantee that effects of treatments remain unchanged beyond the study. Here, we illustrate the feasibility, benefits and cost-effectiveness of enriching standard trial design with electronic follow up. We completed a 5-year electronic follow up of a RCT investigating the impact of probiotics on asthma and eczema in children born 2005–2007, with traditional fieldwork follow up to two years. Participants and trial outcomes were identified and analysed after five years using secure, routine, anonymised, person-based electronic health service databanks. At two years, we identified 93% of participants and compared fieldwork with electronic health records, highlighting areas of agreement and disagreement. Retention of children from lower socio-economic groups was improved, reducing volunteer bias. At 5 years we identified a reduced 82% of participants. These data allowed the trial's first robust analysis of asthma endpoints. We found no indication that probiotic supplementation to pregnant mothers and infants protected against asthma or eczema at 5 years. Continued longer-term follow up is technically straightforward.

Follow up of trial participants on routine electronic health care databases offers great potential to maximise the economic efficiency of trials and allow access to a fuller range of study outcomes and potential adverse events[1–4]. Thus, the future use of electronic databases in clinical trials has been heralded as one of the major benefits to follow construction of nationwide electronic health records systems. To date, however, few studies have demonstrated this benefit, or formally assessed the relationship between traditional fieldwork data and electronic health care databases. Here, we report on the feasibility and efficiency of electronic follow up, and compare traditional trial fieldwork follow up with electronic follow up. We show new insights gained from outcomes electronically recorded 3 years after the end of the trial, and the congruence between fieldwork and electronic data.

Methods

The PROBAT Trial. The randomised, double-blind, placebo-controlled, parallel-group trial on which this follow-up is based was an assessment of the impact of a probiotic food supplement on prevention of the atopic conditions asthma, and eczema in young children. Healthy pregnant women were recruited between 2005 and 2007 to take supplement, or placebo, from 36 weeks' gestation, and then administer it to their infants during their first 6 months of life. The supplement consisted of *lactobacilli* (*Lactobacillus salivarius* CUL61 NCIMB 30211 and *Lactobacillus paracasei* CUL08 NCIMB 30154) and *bifidobacteria* (*Bifidobacterium animalis subsp. lactis* CUL34 NCIMB 30172 and *Bifidobacterium bifidum* CUL20 NCIMB 30153) strains, administered at 1 × 10^{10} cfu per day[5,6]. The trial preparations were not marketed, but it remains a limitation that children or breastfeeding mothers in both arms may have taken commercial probiotics or live yoghurts during the extended follow up,

[1]Swansea University Medical School, Singleton Park, Swansea, UK. [2]Department of Nursing, The College of Human and Health Sciences, Swansea University, Singleton Park, Swansea, UK. [3]The Children's Trust, Tadworth, Surrey, UK. [4]The Harley Street Clinic Children's Hospital, London, UK. [5]Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, UK. [6]Research Department, Cultech Limited, Baglan Industrial Park, Port Talbot, UK. Gareth Davies, Sue Jordan and Mike Gravenor jointly supervised this work. Correspondence and requests for materials should be addressed to S.J. (email: S.E.Jordan@swansea.ac.uk)

however our primary hypotheses were based on elucidating the impact of very early exposure to specified organisms. Participants completed researcher-administered questionnaires, and attended clinics at regular intervals. Information was sought on health status, risk factors and symptoms of atopy, and adverse events[5]. Outcomes included associations between the probiotic supplement and reports of eczema and asthma at or before 2 years of age (see Tables S1, S2 in Supplementary Material).

Linking trial and health service electronic data. The trial location allowed access to person-based routinely-collected electronic data collected by the National Health Service (NHS) and other public sector organisations in the UK, maintained at the Secure Anonymised Information Linkage (SAIL) databank at Swansea University Medical School[7,8]. Following ethical approval from South West Wales Research Ethics Committee on behalf of NHS Wales (project ref. 2004024), participants gave signed informed consent to 5-year follow up through medical records. This randomised, double blind, parallel group, placebo controlled trial is registered with Current Controlled Trials ISRCTN 26287422. The medical information accessed from General Practice (GP) and primary care records included all prescriptions and diagnoses. SAIL links a wide range of person-based datasets using robust irrevocable anonymisation, offering a valuable research resource whilst complying with UK data protection legislation[9] and ensuring confidentiality. Following Information Governance Review Panel approval (reference number 0028), data were acquired from the databank in 2013 (after the last participant was 5 years old), and analysed 2014–15. All methods were performed in accordance with the terms and conditions of the Research Ethics Committee and Governance approvals (above).

Unique personal identifiers in PROBAT were trial study numbers linked to original recruitment records. Within SAIL, the identifier is known as an anonymised linkage field (ALF). The ALF is encrypted by a trusted third party [NHS Wales Informatics Service] and linked to clinical records within SAIL[10]. These two pieces of information were combined thus uploading the fieldwork into SAIL alongside health service data. In linking the data sets it was straightforward to retain blinding of study groups for all analysis, an advantage over open label study follow up. Confidentiality is maintained by the trusted third party anonymisation process, which includes reducing birthdates to week of birth and address to local area units of approximately 1500 people and if necessary aggregation of data across individuals. The area information allows calculation of Townsend indices of social deprivation. Data cannot be extracted from SAIL if patient anonymity can be compromised[10], and published summary statistics must not identify a group of individuals less than 5 in number. Once all individual data is analysed in SAIL, odds ratios may need to be approximated (smoothed to a lower number of decimal points or significant figures) to ensure this. In our study, this was a very minor issue. On occasions we needed to restrict the odds ratio precision to 1 decimal point; these are indicated in the tables as 'approximate' statistics.

SAIL data were examined for NHS 5-byte version 2 Read codes[11] relating to diagnoses of asthma and eczema, asthma related medicines, and antibacterial prescriptions. Asthma diagnosis in childhood is complex, and becomes more reliable as a child ages. As a primary outcome definition, rather than simply reported (Read coded) diagnoses at this young age (2–5 years); we accessed and included prescribed medication, specifically: more than one prescription of a beta2 agonist, or, a recorded diagnosis (Read code) of asthma plus at least one single prescription of a beta2 agonist. For sensitivity analysis, we also considered a range of 5 definitions, representing increased likelihood of a true positive diagnosis, classified according to British Thoracic Society guidelines[12,13] (Tables S3, S4). Asthma definitions were run as Structured Query Language (SQL) queries based on lists of Read codes for (anonymised) trial participants in the databank. Eczema diagnosis was relatively straightforward, defined as presence of Read codes for diagnoses of infantile eczema, flexural eczema, infected eczema, atopic dermatitis/eczema, and eczema not otherwise specified.

A crucial factor is the extent of electronic data coverage. In the setting of this trial, GP practices are incentivised to accurately code certain medical conditions under the Quality and Outcomes Framework[14,15], a process of financial reward based on points systems for certain clinical indicators, including long term monitoring of patients with asthma. Like many study resources, electronic records have missing information. However, unlike a designed study, one cannot easily identify missing cells. A gap in the electronic record can reflect failure of the record, or simply absence of any clinical event at that time. We defined 'Good' data coverage, and therefore successful follow up, according to an algorithm based on: the frequency of any event in the GP practice in the period being at least 10% of that measured from Jan 1st to Dec 30th 2009 (a period known to have good coverage); the absence of gaps in the individual's or GP practice's record of greater than 30 days. It is noted that the algorithm is *ad hoc*, but serves to identify GP practices that do not submit data regularly and individuals who have deregistered from their primary care practices, for example by moving away from the area.

Statistical Analysis. For children (trial subjects) present with good data coverage at 2 and 5 years, trial outcomes and covariates were assessed using binary logistic regression. Any variables identified as statistically significant (5% level) using univariable analysis were investigated in multivariable logistic regression adjusted models, with a final model chosen by likelihood ratio tests (with a variable retained in the final model if the model showed a significant improvement in fit). To explore the sensitivity of the conclusions to issues of trial compliance and uncertainty in the definition of asthma we undertook ITT (intention to treat) and PP (*per protocol*) analyses for asthma and eczema at 5 years, in combination with all 5 reasonable asthma classifications. We defined good compliance as infants taking >70% of the treatment (30 or more treatments) by 6 weeks of age.

Comparison of electronic and field data was made using Cohen's Kappa, with levels of agreement as defined by Landis *et al.*[16]. In all tables we report p-values without adjustment for multiple comparisons. This was due to the high degree of correlation between endpoints (for example the different, but nested, definitions of asthma) making formal adjustment difficult. However we note that without any adjustment, P-values should be treated with the usual caution, and interpreted within the context of the number of tests performed. All analyses were completed before treatment group allocation was un-blinded.

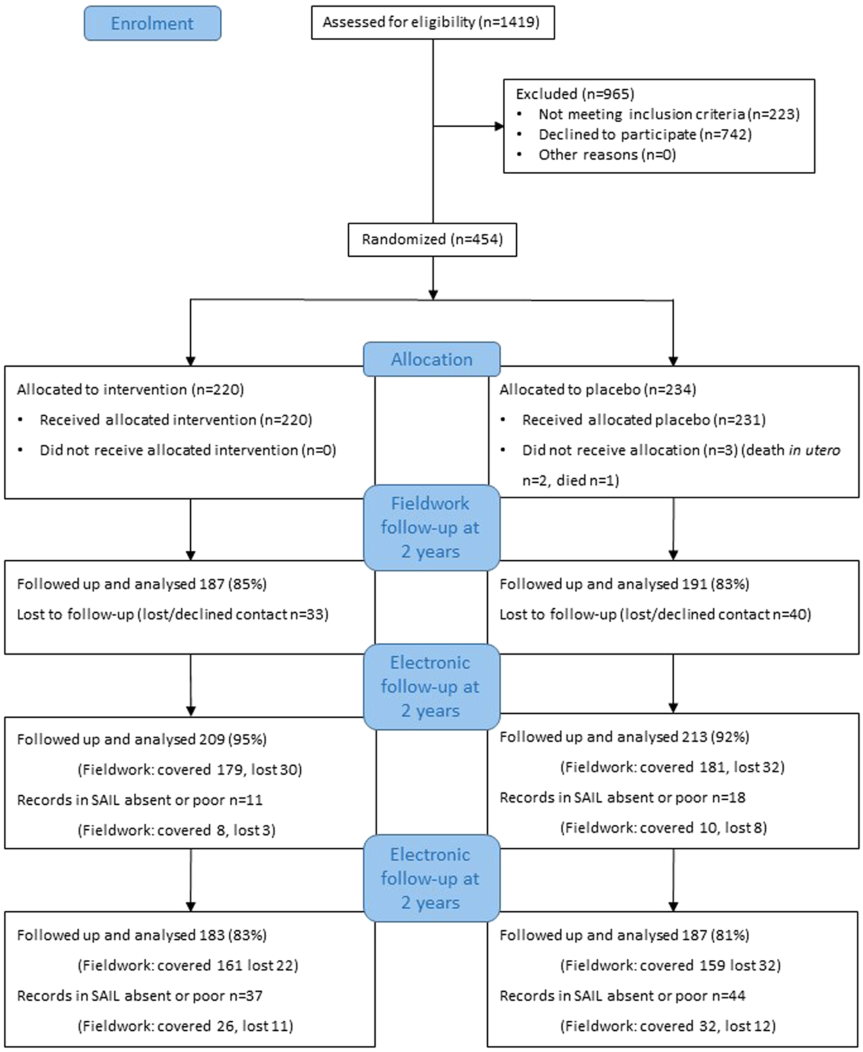
Enrolment
Assessed for eligibility (n=1419)
Excluded (n=965)
• Not meeting inclusion criteria (n=223)
• Declined to participate (n=742)
• Other reasons (n=0)
Randomized (n=454)
Allocation
Allocated to intervention (n=220)
• Received allocated intervention (n=220)
• Did not receive allocated intervention (n=0)
Allocated to placebo (n=234)
• Received allocated placebo (n=231)
• Did not receive allocation (n=3) (death *in utero* n=2, died n=1)
Fieldwork follow-up at 2 years
Followed up and analysed 187 (85%)
Lost to follow-up (lost/declined contact n=33)
Followed up and analysed 191 (83%)
Lost to follow-up (lost/declined contact n=40)
Electronic follow-up at 2 years
Followed up and analysed 209 (95%)
(Fieldwork: covered 179, lost 30)
Records in SAIL absent or poor n=11
(Fieldwork: covered 8, lost 3)
Followed up and analysed 213 (92%)
(Fieldwork: covered 181, lost 32)
Records in SAIL absent or poor n=18
(Fieldwork: covered 10, lost 8)
Electronic follow-up at 2 years
Followed up and analysed 183 (83%)
(Fieldwork: covered 161 lost 22)
Records in SAIL absent or poor n=37
(Fieldwork: covered 26, lost 11)
Followed up and analysed 187 (81%)
(Fieldwork: covered 159 lost 32)
Records in SAIL absent or poor n=44
(Fieldwork: covered 32, lost 12)

Figure 1. Trial Profile: the PROBAT trial with 5 year follow up. see refs[5,6,40].

Results

Feasibility of electronic follow-up.

435 of 452 (96%) trial participants were located in the SAIL databank. Good coverage was identified for 422 participants at 2 years (93%) and 370 participants at 5 years (82%). The electronically extended trial profile is illustrated in Fig. 1. Interestingly, loss-to-follow up of the most deprived (assessed by Townsend score) was less in electronic than traditional trial data. Figure 2 shows that although great efforts were initially made to start with a representative proportion of the lowest socio-economic fifth, this group was more likely to be lost to follow up in field work. In contrast, these participants were recovered in 2 year electronic follow up and retained at 5 year electronic follow up (see also Table S1 for demographic details).

Comparison of electronic data with fieldwork at 2 years.

For asthma outcomes, 318 participants were present in both electronic and fieldwork datasets. For eczema and antibacterial usage, the numbers were 321 and 408 respectively. All comparisons between electronic and field data showed 'fair' agreement (kappa values 0.26 to 0.54), but all also fell short of 'substantial' agreement (a kappa value >0.60) (Table S9).

The disagreements can be examined in more detail. 23 participants were classified as having asthma in the 2 year fieldwork. Of these, only 10 had a general practitioner diagnosis or had been prescribed >1 beta2 agonist prescription in the electronic records. Coincidentally, exactly 23 participants were also classified as asthmatic in the electronic data, with 10 of these not having been reported as such to the fieldworkers. Agreement for antibiotic exposure was little better than chance.

Analysis of trial outcomes using electronic follow-up data at 2 and 5 years.

Briefly, at 2 and 5 years, both ITT and PP analyses indicated a higher prevalence of asthma in the children in the probiotic arm in univariable and multivariable analyses. This difference was statistically significant (not adjusted for multiple comparisons) in PP analyses only, and in both sets of analyses the association at 5 years was less than that found at 2 years. Table 1 shows absolute risk differences for the treatment groups at 5 year electronic follow up. The number of participants remaining in the PP analyses was considerably reduced (Table 1), which was a consequence of a

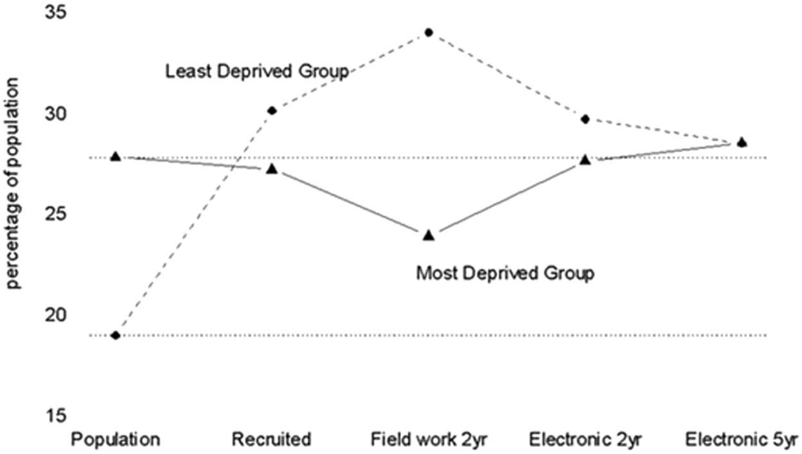

Figure 2. Trends in recruitment and retention of trial participants via fielder ork and electronic follow up. Our study was conducted in a population with a high percentage (28%. dotted horizontal line) of people in the lowest socio-economic fifth (defined by the Townsend Index), and low percentage in the highest socio-economic fifth (18%). Although recruitment to the trial from the most deprived group was excellent, this group was under-represented in field work follow-up at 2 years. The impact of this was explored elsewhere[3]. However, this loss was rectified in electronic records at both 2 years and 5 years (solid line). The trial over-recruited in the least deprived group; this discrepancy was exacerbated in field work, but improved by electronic follow up (dashed line). Deprivation (Townsend) fifths are based on geographical area of residence, using Lower Super Output Areas (LSOAs) defined by postcodes. This measure of material deprivation is calculated from rates of unemployment, vehicle ownership, home ownership, and overcrowding[45].

	n(%) with condition exposed to probiotics	n(%) with condition not exposed	Absolute risk difference (95% CI) Between exposed and not exposed	NNH
Asthma at 5 years *ITT analysis*	43/183 (23.5%)	40/187 (21.4%)	2.1% (−6.4–10.6%)‡	48
Asthma at 5 years *PP* analysis	26/84 (31.0%)	15/88 (17.0%)	13.9% (1.3–7.9%)‡ $p < 0.05$	8
				NNT
Eczema 5 years *ITT analysis*	61/183 (33.3%)	65/187 (34.8%)	1.4% (−8.2 to 11.1%)$^\$$	71
Eczema 5 years *PP* analysis	29/84 (34.5%)	33/88 (37.5%)	3.0% (−11.4 to 17.3%)$^\$$	34

Table 1. Main asthma and eczema outcomes at 5 years. Asthma primary definition based on prescription of more than one prescription of a beta2 agonist, or, a recorded diagnosis of asthma plus at least one single prescription of a beta2 agonist. ‡Favours placebo. $^\$$Favours probiotics. NNH = number needed to harm, NNT = number needed to treat. Full details are in Supplementary Tables 5–8 and, for 2 year follow up, in ref.[6], Table 5.

strict definition of compliance (>30 treatments), providing a clear contrast with the ITT groups. Approximate (to comply with SAIL data release policy[10]) and exact unadjusted odds ratios were: 2 yr PP OR ~5 (1.6 to 15.0); 5 yr PP OR = 2.18 (1.06 to 4.50); 2 year ITT OR = 1.43 (0.74 to 2.75); 5 yr ITT OR = 1.13 (0.69 to 1.84).

We next explored multivariable analysis, adjusting for background variables. In PP analysis at 2 years, no other demographic variables were associated with asthma. In PP analysis at 5 years, the probiotic arm was retained as significant in the final model, after adjustment for significant risk factors: smoking and paternal history of asthma. The odds ratio increased to 3.33 (1.47 to 7.14). Sensitivity of results to the definition of asthma, and compliance is given in full in Table S10. We found, after adjustment for any significant background variable, that the estimated odds ratio for the effect of the probiotic on asthma ranged between a high of 5.0 (asthma definition 2 at 2 years, PP analysis, 95% CI (1.5, 15) p = 0.01) to a low of 0.41 (asthma definition 3 at 2 years, ITT analysis, 95% CI (0.14, 1.2) p = 0.1). One odds ratio was exactly 1.0; 9/17 showed an odds ratio greater than one, and 7/17 showed an odds ratio below one. Thus there was no consistent evidence for a probiotic effect. Significant background risk factors included family history of allergy, absence of breastfeeding and smoking in the household (S10).

In all treatment group analyses, odds ratios for eczema were less than 1, indicating a trend for lower eczema in the probiotic arm, but no comparisons reached statistical significance at the 5% level. Demographic risk factors associated with eczema included attendance at daycare, keeping caged birds or rodents as pets, and paternal reported allergy. Full details are available in Tables 5–8 and sensitivity to compliance table in S10.

Discussion

The benefit of electronic record linkage to long term trial follow was shown by the West of Scotland Coronary Prevention Study[1,17], which was instrumental in expanding statin use for cardiovascular disease prevention. 10 years after the end of the trial, continued risk reduction was demonstrated along with no long-term safety concerns (see also the follow-up of the ISIS-2 trial via hospital records[18]). However, these approaches are still neither routine practice nor requirements of funding bodies. Using keywords 'electronic health record', 'data linkage', 'clinical trial', 'follow-up', 'secondary use', 'comparative study', in PubMed and SCOPUS (21/08/17) we found no other studies comparing findings from electronic health services' databases with traditional fieldwork. This despite great concerns over study replication, a burgeoning literature proposing the use of electronic medical records in clinical research[19–23] and projects such as The Electronic Healthcare Records for Clinical Research[24]. Here we show electronic follow up of an RCT, both at 2 years (within the trial period) and 5 years (beyond available resources in the traditional trial framework). We also introduce a further level of linkage: pregnant women were recruited to the trial, but follow up was extended to the more relevant population; children identified in the health records.

The proportion of recruits withdrawing from trials steadily increased between 1955 and 2000[25], prompting the US Food and Drug Administration to insist on measures to minimise missing data[26]. This is not simply due to loss of power, but also the consequent bias if loss is non-random[27]. Where participants are lost from the socioeconomic group where disease prevalence is highest, usually the most deprived, power is further reduced[6]. Indeed, in our trial the lowest socio-economic group was less represented, but when electronic records were used, volunteer selection bias decreased. We note, given our loss to follow up is 18% at 5 years, that additional biases related to inclusion in databases may still be present. However, this is within the 80–90% follow up rates that are generally considered acceptable[28].

Electronic medical records are increasingly used in observational studies to detect safety issues, such as adverse drug events[29] and prescription errors[30]. SAIL has been used as a cost-effective resource for trial recruitment[31], and evaluation of complementary and alternative medicines[32]. In SAIL cohort studies (ankylosing spondylitis[33], psychotic disorders[34]), 80% of patients were located within the electronic records, comparable figures to our study at 5 years.

The level of agreement between data sources is difficult to interpret, especially where fieldwork involves reliance on carer recall. Neither fieldwork nor the electronic record can be considered the gold standard. Some endpoints are better suited to traditional fieldwork, and others to electronic health records. Fieldwork data are vulnerable to misunderstandings of questionnaires, perceptions or definitions of illness. Although prescription recording is accurate, the prescription itself may not be redeemed or used by the carer (see Gadkari *et al.*)[35]. Asthma has been identified as a condition for which compliance with primary care prescription is low[36]. Here, we propose the asthma endpoints are generally more reliable in the electronic follow up, due to the older age of the children and also the risk of bias or unreliability in carer-reported outcomes, where a range of concerns might be designated 'asthma'. This might be reflected in the high proportion of children with fieldwork-reported asthma, yet no evidence of any prescription of asthma medicines. Antibiotic usage is likely a case where the more objective electronic data is preferred, due to difficulty in recalling medicine up to 12 months ago in the fieldwork, or misunderstandings over the term. On the other hand, electronic data may also be compromised by in-hospital prescription not recorded in the GP data. In contrast, certain outcomes such as atopic eczema cannot easily be reproduced in the electronic follow up, as the link to prescription medicines is less clear.

Impact of probiotics on Clinical Outcomes. The electronic data at 2 years were in agreement with the original trial, and other literature, in failing to find a preventive role for probiotics on asthma outcomes[37,38]. We identified a possible increased asthma prevalence[39,40], though this should be interpreted with caution as several endpoints were analysed in sensitivity analyses. Our study is the first to address longer term electronic follow up of infants exposed *in utero* and suggests that even if there are any associations, they tend to diminish over time.

In the original trial, at 2 years, the probiotic supplementation appeared to confer some benefits by reducing atopic eczema (not 'any eczema')[6,41]. This diagnosis required skin prick testing and was therefore unsuited to electronic follow up. Meta-analysis[42] of studies investigating probiotics' impact on more widely defined eczema, as investigated here, suggests a possible preventative effect. This was not supported by any statistically significant findings in our study. We note that our estimated effect sizes were consistent with published pooled relative risks of between 0.62 and 0.78, however we also found that the odds ratio in our study moved closer to 1 between 2 and 5 years follow up.

Strengths and Limitations. The advantages of electronic follow up include minimisation of reporter, recall, social desirability response bias or mis-understandings in questionnaires. However, constraints or missing values in health records may include: absence of medicines prescribed in secondary care (soon to be available in SAIL) or privately purchased; absence of free text clinical notes; prescriptions given 'just in case'; non-adherence to prescription regimens; quality of coding; absence of confounding variables (although many demographic variables can be recorded, trials will often require additional information). We acknowledge that prescriptions of asthma medicines at 2 and 5 years may not indicate a formal diagnosis of asthma; rather, they are functional measures of primary care practitioners' perceptions of respiratory illness. Extension to beyond 7 years is straightforward (with ethical approval) and would offer a definitive asthma endpoint and further illustration of the advantages of the enriched trial.

Electronic follow up is likely to be cost-effective. Health informaticians can match and extract data quickly, here, some 6 months of analyst's time, plus statistical re-analysis, research governance approvals, checking against data held by the Office of National Statistics (ONS), and checking with fieldwork data. We estimate the cost to be $80k ($57k in human resources), ~$200 per participant, which compares favourably with a mean of $24,727 per participant for patient retention in phase 3 pharmaceutical trials[43]. Although full cost-benefit analysis would require quantification of data quality, unlike traditional methods, the cost of follow-up using routine data is potentially relatively small and crucially does not increase with the number of participants. However, an infrastructure must be in place and meet exacting safeguarding standards[44]. SAIL, for example, is a large project supported by data scientists and significant computing hardware. This approach is still cost-effective over even a small number of trials. Given the cost of a suite of trials (our experience with PROBAT was approximately $2000 per participant) it is imperative that as much information, over a period as long as possible, is utilised. The potential bias introduced whenever there is loss of data within the trial framework must always be considered, however we suggest that the benefits of updating and re-testing trial hypotheses are such that trial funders should routinely recognise and support infrastructure for long-term follow up, using centres offering comprehensive electronic follow up[7,8].

Key Points.

- Electronic follow-up of a clinical trial involving UK participants using the SAIL databank was feasible and effective, with a high proportion (96%) of participants identified with good data coverage.
- At 2 years of age retention was greater in electronic records, and with reduced attrition bias in lower socio-economic status groups.
- 5 years, retention was still high and free of bias in socio-economic status. Future extension of the trial is straightforward within this framework.
- 5 year electronic follow up did not find support for an effect of early use of probiotics on childhood eczema or asthma.
- The benefits relating to cost-effective, long term monitoring of complex interventions have implications for future clinical trial design.

References

1. Ford, I. *et al*. Long-term follow-up of the West of Scotland Coronary Prevention Study. *N Engl J Med* **357**, 1477–1486, https://doi.org/10.1056/NEJMoa065994 (2007).
2. Berry, J. G. *et al*. A randomised controlled trial to compare opt-in and opt-out parental consent for childhood vaccine safety surveillance using data linkage: study protocol. *Trials* **12**, 1, https://doi.org/10.1186/1745-6215-12-1 (2011).
3. Calver, J. *et al*. Does health assessment improve health outcomes in indigenous people? An RCT with 13 years of follow-up. *Aust N Z J Public Health* **29**, 107–111 (2005).
4. Holman, C. D. *et al*. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Aust Health Rev* **32**, 766–777 (2008).
5. Allen, S. J. *et al*. Dietary supplementation with lactobacilli and bifidobacteria is well tolerated and not associated with adverse events during late pregnancy and early infancy. *J Nutr* **140**, 483–488, https://doi.org/10.3945/jn.109.117093 (2010).
6. Jordan, S. *et al*. Volunteer bias in recruitment, retention, and blood sample donation in a randomised controlled trial involving mothers and their children at six months and two years: a longitudinal analysis. *PLoS One* **8**, e67912, https://doi.org/10.1371/journal.pone.0067912 (2013).
7. Ford, D. V. *et al*. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* **9**, 157, https://doi.org/10.1186/1472-6963-9-157 (2009).
8. Lyons, R. A. *et al*. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* **9**, 3, https://doi.org/10.1186/1472-6947-9-3 (2009).
9. Gov.uk. (2015). Data protection - GOV.UK. Available at: https://www.gov.uk/data-protection/the-data-protection-act. Accessed 7th Sept. 2016.
10. Swansea University. Operating Model. SAIL. www.saildatabank.co.uk/operating-model Accessed May 12, 2016 (2014).
11. Health and Social Care Information Centre. Read Codes. http://systems.hscic.gov.uk/data/uktc/readcodes. AccessedMay 12, 2016.
12. British Thoracic Society, Scottish Intercollegiate Guidelines Network. British Guideline on the Management of Asthma: A national clinical guideline. London: *British Thoracic Society* (2012).
13. British Thoracic Society, Scottish Intercollegiate Guidelines Network. British Guideline on the Management of Asthma: A national clinical guideline. Edinburgh: *Healthcare Improvement Scotland* (2014).
14. Roland, M. Linking physicians' pay to the quality of care–a major experiment in the United kingdom. *N Engl J Med* **351**, 1448–1454, https://doi.org/10.1056/NEJMhpr041294 (2004).
15. Doran, T. *et al*. Pay-for-performance programs in family practices in the United Kingdom. *N Engl J Med* **355**, 375–384, https://doi.org/10.1056/NEJMsa055505 (2006).
16. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
17. Barry, S. J., Dinnett, E., Kean, S., Gaw, A. & Ford, I. Are routinely collected NHS administrative records suitable for endpoint identification in clinical trials? Evidence from the West of Scotland Coronary Prevention Study. *PLoS One* **8**, e75379, https://doi.org/10.1371/journal.pone.0075379 (2013).
18. Baigent, C. *et al*. ISIS-2: 10 year survival among patients with suspected acute myocardial infarction in randomised comparison of intravenous streptokinase, oral aspirin, both, or neither. The ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. *BMJ* **316**, 1337–1343 (1998).
19. van Velthoven, M. H., Mastellos, N., Majeed, A., O'Donoghue, J. & Car, J. Feasibility of extracting data from electronic medical records for research: an international comparative study. *BMC Med Inform Decis Mak* **16**, 90, https://doi.org/10.1186/s12911-016-0332-1 (2016).
20. Staa, T. P. *et al*. Pragmatic randomised trials using routine electronic health records: putting them to the test. *BMJ* **344**, e55 (2012).
21. Elkhenini, H. F. *et al*. Using an electronic medical record (EMR) to conduct clinical trials: Salford Lung Study feasibility. *BMC Med Inform Decis Mak* **15**, 8, https://doi.org/10.1186/s12911-015-0132-z (2015).
22. Nicholls, S. G. *et al*. The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement: Methods for Arriving at Consensus and Developing Reporting Guidelines. *PLoS One* **10**, e0125620, https://doi.org/10.1371/journal.pone.0125620 (2015).
23. Hruby, G. W., Matsoukas, K., Cimino, J. J. & Weng, C. Facilitating biomedical researchers' interrogation of electronic health record data: Ideas from outside of biomedical informatics. *J Biomed Inform* **60**, 376–384, https://doi.org/10.1016/j.jbi.2016.03.004 (2016).
24. Doods, J. *et al*. Piloting the EHR4CR feasibility platform across Europe. *Methods Inf Med* **53**, 264–268, https://doi.org/10.3414/ME13-01-0134 (2014).
25. Wahlbeck, K., Tuunainen, A., Ahokas, A. & Leucht, S. Dropout rates in randomised antipsychotic drug trials. *Psychopharmacology (Berl)* **155**, 230–233 (2001).
26. O'Neill, R. T. & Temple, R. The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. *Clin Pharmacol Ther* **91**, 550–554, https://doi.org/10.1038/clpt.2011.340 (2012).

27. S. J. & G. M. Vol. 35 36–40 (The Welsh Paediatric Journal, 2011).
28. Fewtrell, M. S. *et al.* How much loss to follow-up is acceptable in long-term randomised trials and prospective studies? *Arch Dis Child* **93**, 458–461, https://doi.org/10.1136/adc.2007.127316 (2008).
29. Linder, J. A. *et al.* Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting. *Pharmacoepidemiol Drug Saf* **19**, 1211–1215 (2010).
30. Abramson, E. L. *et al.* A long-term follow-up evaluation of electronic health record prescribing safety. *J Am Med Inform Assoc* **20**, e52–58, https://doi.org/10.1136/amiajnl-2012-001328 (2013).
31. Brooks, C. J. *et al.* Use of a patient linked data warehouse to facilitate diabetes trial recruitment from primary care. *Prim Care Diabetes* **3**, 245–248, https://doi.org/10.1016/j.pcd.2009.06.004 (2009).
32. Brophy, S. *et al.* Internet-based randomised controlled trials for the evaluation of complementary and alternative medicines: probiotics in spondyloarthropathy. *BMC Musculoskelet Disord* **9**, 4, https://doi.org/10.1186/1471-2474-9-4 (2008).
33. Husain, M. J. *et al.* HERALD (health economics using routine anonymised linked data). *BMC Med Inform Decis Mak* **12**, 24, https://doi.org/10.1186/1472-6947-12-24 (2012).
34. Economou, A. *et al.* The health informatics cohort enhancement project (HICE): using routinely collected primary care data to identify people with a lifetime diagnosis of psychotic disorder. *BMC Res Notes* **5**, 95, https://doi.org/10.1186/1756-0500-5-95 (2012).
35. Gadkari, A. S. & McHorney, C. A. Medication nonfulfillment rates and reasons: narrative systematic review. *Curr Med Res Opin* **26**, 683–705, https://doi.org/10.1185/03007990903550586 (2010).
36. Gutiérrez, S. J., Fagnano, M., Wiesenthal, E., Koehler, A. D. & Halterman, J. S. Discrepancies between medical record data and parent reported use of preventive asthma medications. *J Asthma* **51**, 446–450, https://doi.org/10.3109/02770903.2013.878351 (2014).
37. Tang, L. J., Chen, J. & Shen, Y. Meta-analysis of probiotics preventing allergic diseases in infants. *Zhonghua Er Ke Za Zhi* **50**, 504–509 (2012).
38. Fiocchi, A. *et al.* Clinical Use of Probiotics in Pediatric Allergy (CUPPA): A World AllergyOrganization Position Paper. *World Allergy Organ J* **5**, 148–167, https://doi.org/10.1097/WOX.0b013e3182784ee0 (2012).
39. Kopp, M. V., Hennemuth, I., Heinzmann, A. & Urbanek, R. Randomized, double-blind, placebo-controlled trial of probiotics for primary prevention: no clinical effects of Lactobacillus GG supplementation. *Pediatrics* **121**, e850–856, https://doi.org/10.1542/peds.2007-1492 (2008).
40. Kalliomäki, M., Salminen, S., Poussa, T., Arvilommi, H. & Isolauri, E. Probiotics and prevention of atopic disease: 4-year follow-up of a randomised placebo-controlled trial. *Lancet* **361**, 1869–1871, https://doi.org/10.1016/S0140-6736(03)13490-3 (2003).
41. Allen, S. J. *et al.* Probiotics in the prevention of eczema: a randomised controlled trial. *Arch Dis Child* **99**, 1014–1019, https://doi.org/10.1136/archdischild-2013-305799 (2014).
42. Dang, D. *et al.* Meta-analysis of probiotics and/or prebiotics for the prevention of eczema. *J Int Med Res* **41**, 1426–1436, https://doi.org/10.1177/0300060513493692 (2013).
43. Sertkaya, A., Wong, H. H., Jessup, A. & Beleche, T. Key cost drivers of pharmaceutical clinical trials in the United States. *Clin Trials* **13**, 117–126, https://doi.org/10.1177/1740774515625964 (2016).
44. U.S. Department of Health and Human Services. Food and Drug Administration. Guidance for Industry: Electronic Source Data in Clinical Investigations. *Silver Springs, MD: U.S. Department of Health and Human Services. Food and Drug Administration* (2013).
45. Townsend, P., Phillimore, P. & Beattie, A. *Health and deprivation: inequality and the north*. (Croom Helm, 1988).

Acknowledgements

This randomised, double blind, parallel group, placebo controlled trial is registered with Current Controlled Trials ISRCTN 26287422: registered title: Probiotics in the prevention of atopy in infants and children. Ethical approval was granted in February 2004 by the South West Wales Research Ethics Committee on behalf of NHS Wales (project ref. 2004024). The trial was supported by the Knowledge Exploitation Fund, Collaborative Industrial Research (Project number: HE09 COL 1002), Welsh Development Agency, UK and Cultech Limited, Unit 3, Christchurch Rd, Baglan Industrial Park, Port Talbot, SA12 7BZ, UK. This work is part-funded by the European Social Fund (ESF) through the European Union's Convergence programme administered by the Welsh Government. Cultech Ltd. UK part funded the trial and provided the probiotic and matching placebo, and generated the random allocation sequence. The follow up study was funded by the Knowledge Economy Skills Scholarships (KESS). Approval for the follow up study was granted by the HIRU Information Governance Review Panel (IGRP) on 29th January 2010 (reference number 0028). Sue Plummer is a Director of Cultech and advised on study design and contributed to the final report. The sponsors were not involved in data collection, analysis or interpretation of the findings. The Knowledge Exploitation Fund had no involvement in the conduct of the trial. Sue Jordan had full access to all the data in the study and had final responsibility for the decision to submit for publication. This study uses anonymised data held in the Secure Anonymised Information Linkage (SAIL) system, which is part of the national e-health records research infrastructure for Wales (http://www.saildatabank.com/). Data held in SAIL databases are anonymised and aggregated and have been obtained with permission of relevant Data Protection Officers, as approved by the National Research Ethics Service, Wales. We should like to acknowledge all the data providers who make anonymised data available for research. Thanks are due to: our participants; colleagues in Abertawe Bro Morgannwg University Health Board, particularly Vivienne Davies, phlebotomist, Allyson James, clinic nurse; colleagues in Swansea University, particularly research assistants and data managers Sally Williams, Julia Kramer, Amanda Cook, Ceri Bradshaw, Ioan Humphries, and Claire Burrows.

Author Contributions

G.D. contributed to the project writeup and data analysis, S.J. and M.G. contributed to study design and project writeup. Support with structured query language data queries was provided by C.B. and D.T. D.T. also provided the data cleaning algorithm to assess good coverage. M.S. contributed with referencing and sample collection for the fieldwork trial. G.M. and S.A. provided clinical advice regarding asthma in children. I.G. contributed towards study design and advice regarding probiotics. S.P. provided funding for the fieldwork trial, help with study design and advice regarding probiotics.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-25954-z.

Competing Interests: S.J. Allen, S. Jordan, M.B. Gravenor, D. Thayer, C. Brooks and G. Morgan, no conflicts of interest. G. Davies and M. Storey received financial support from Cultech Ltd. At the time the study was undertaken, I.G. and S.F.P. were employees of Obsidian Research Ltd, and are now employees of Cultech Ltd.