Charity and CIC match-checking: enhanced file January 2021

Abstract

This document describes a new file which extends and entirely supersedes the previous one (December 2020). The new file is named

Charity and CIC checking Results v2

with various appropriate filename extensions.

The previous file contained the combined results of the checking (and where necessary amendment) of automatic matches to official registers of suppliers contained in three original files, one comprising 1806 records believed to represent charities, the second 241 records believed to be CICs, and the third a further 43 possible charities.

That dataset has now been significantly extended with further, previously unrecognised, Third Sector suppliers found within the wider database. The addition and processing of those suppliers has also led, in a small number of cases, to improvements to the matching of suppliers in the three original files.

Further, improvements to the wider database have also permitted the elimination of a number of records from the previous file, which had been necessary to permit the correct matching of suppliers whose names had become corrupted to the point of containing all or parts of the names of two 'true' suppliers. The main sources of those corruptions in the source data have been resolved, eliminating the corrupted supplier names from the database, and thereby eliminating the need to include them in the present file.

The result of the additional work is that the present file contains 1820 suppliers identified as Registered Charities, 1776 of them incorporated (up from 1620 and 1598), and 295 CICs (up from 241), plus 6 organisations that have been both Charity and CIC at times during or since the study period. Just 23 of the charities (of which 18 are unincorporated) are explicitly associated with NHS entities - typically, hospitals' charitable funds.

This covering document has also been significantly updated to accord with the augmented data file, and thus also entirely supersedes its previous version.

Contents

1. Introduction

This brief document concentrates on describing the file containing the results of the checking and correction of matches; discussion of the process itself is restricted only to what is necessary for that purpose. The file has been considerably enhanced and augmented since the previous (December 2020) issue. The changes, in decreasing sequence of number of records affected, are summarised below.

1.1 Improvements since previous file issue

More Third Sector suppliers identified and included The files that started this process resulted from an automatic matching process that had been developed and tuned to minimise the number of false matches it made, but also to balance them: to avoid bias by equating approximately the numbers of 'false positive' and 'false negative' matches. The 'false positives', of course, were included in the original files of Charities and CICs provided for checking; but logically there would remain elsewhere in the wider dataset the 'false negatives'. On that presumption of balance, and given that the number of false positives had been manageable for manual checking, steps were taken to find the false negatives too - to cast the net wider, in effect. This was achieved in a number of ways, initially including:

- Identifying suppliers with names that were 'similar' to those of suppliers already identified as Third Sector;
- Identifying suppliers with names that were 'similar' to names (formal, working or previous) of registered Charities (the closest matches had been included in the original file of course);
- Identifying suppliers with names containing any one of a list of several dozen keywords and -phrases that had been observed among the already-matched Third Sector suppliers, and a number of others suggested by experience;
- Including a number of heavily-abbreviated and acronym supplier names that had eluded the auto-matcher.

This yielded a set of just over 400 further suppliers for checking; about half of which turned out to be Third Sector.

More CICs identified and included In an extension of the above process, suppliers were identified which had names 'similar' to those of any registered CIC (whether a supplier or not). This yielded approximately another fifty candidate suppliers, around two-thirds were found upon checking indeed to be CICs.

Elimination of corrupted supplier names The original files, and the additions described above, included a number of 'entangled' supplier names, in which the 'name' actually comprised all or part of two 'true' supplier names. Examination revealed that the vast majority of these arose from just two sources, the transaction data from Surrey Downs CCG and Barnet CCG. The original data from these two CCGs were unusually and inconsistently-formatted, which had resulted in corruption and entanglement when parsed. By different methods, the uniquely-presented Surrey Downs data were re-parsed from scratch, and the supplier names in the Barnet data repaired in situ. This removed the need for over 80 corrupted-name records in the present datafile, which have, accordingly, been discarded, along with a small number of other insoluble supplier name fragments which had been collaterally gathered in the processes described earlier in this section.

More candidates from a late addition to the database At the eleventh hour, data from Central London (Westminster) CCG became available. Within them were a number of suppliers not seen elsewhere in the dataset, ten of which were identified as Third Sector and added to the present file.

Revisions to previous matches One further benefit from the additional records was enhanced insight into certain previously challenging suppliers; this has led to revisions to the matches on a small number of the suppliers that were already in the previous file.

Ad-hoc additions The previous, now-superseded, results file included a small number of manually-added records for Third Sector suppliers that had been spotted by chance in the course of the checking process; the present file also includes a couple more.

The processes described above that 'cast the net wider' inevitably also picked out a number of suppliers that were found, upon checking, not to be Third Sector. These records have nonetheless been retained in the file,

as were those in the original Charities and CICs files that turned out not to be Third Sector. Although such records are of no direct benefit to Third Sector analysis, they nonetheless represent validated matches within the wider file, and perhaps more importantly, contain the information which proves that certain suppliers are not Third Sector, even though their names suggest they might or should be. They thus save future researchers the effort of re-checking suppliers that might otherwise appear to have been overlooked by the present process.

2. Checks

Within the various source files, NHS suppliers are identified only by their name; but that name provides a cross-reference to lists of NHS customers: CCGs, Trusts, and NHS England, in three separate files, and a master list of accepted and rejected 'automatic' matches from a prior process to which all suppliers had been subjected.

The CCG and Trust files identified the NHS customer(s) for each supplier, which in turn identified the part(s) of the country in which the supplier operated, often a vital resource in identifying some of the more tersely-named suppliers. All three source files identified the dates of known transactions (which were relevant in the many cases of suppliers - especially charities - which had changed their legal form and thus registration details during the study period), and the amount paid in each transaction, which could be compared with the income of a supplier's presumptive match(es) to confirm or exclude their viability as potential matches.

In the course of the work it was found that in most cases it was best to try to match the supplier in question directly, based upon its name, location, activities and financial 'size' (as indicated by the number and value of transactions), and then compare that manual result with the 'auto-match'; rather than simply to try to validate the existing match: too many viable, and sometimes preferable, alternatives could be missed that way.

In all cases of new or revised matches, whether as charity, company (including CIC) or both, the registered name and number of the matched organisation were both immediately recorded. When all matching work was complete, these number-and-name pairs were validated by comparison with the relevant regulator's master lists, as proof against typing errors in the recording of manual matches.

In addition to the resulting matches themselves, the file contains categorical variables which indicate the results of the checking process, and where necessary, any uncertainties that remain.

Since time constraints precluded the diligent checking of every supplier in the combined file, various methods were used to identify those most likely to be mismatched. Ultimately, the only matches not manually verified were some of those with the clearest and most specific supplier names, and which had been established in their present form for the longest time: the proportion that have re-registered and/or changed legal form during the past decade is surprisingly high. Cases in which such a change has occurred since the first known NHS transaction are specifically identified in the file.

It was further discovered that a unique supplier name did not necessarily identify a unique supplier, particularly in the case of large organisations with numerous separately-registered local branches: when, for example, CCGs in Shropshire, Durham, Berkshire, and Sussex each recorded a payment to 'Mind', it seems much more likely that each meant its local branch than that all meant the national parent organisation. Such 'multi-supplier' cases are of course also identified and detailed in the file.

3. Results

As received, the original CICs file contained 241 suppliers which appeared to have that legal form, as indicated either by their name or the auto-matching results. The Charities file identified 1806 suppliers, 1717 of which had been auto-matched to charities; the remaining 89 had potential auto-matches to two or more organisations, at least one of which was a charity. The 'extra charities' file contained a further 43 suppliers believed likely to be charities. It was a concern that, of the 1717 auto-matched charities, only 1087 (63%) had also been auto-matched to companies. That is, the results of the auto-matching process seemed to suggest that an unexpectedly large proportion of the registered-charity suppliers were unincorporated: were taking on supply contracts with the NHS without the legal protection of being a registered company.

The results of the match-checking confirmed that 233 of the auto-matched suppliers in the original CIC file had indeed been registered as such for all or part of the study period; a further 62 CICs have been identified in the other sources, and six more that were CICs for part of the study period.

Turning to the charities, of the suppliers in the original Charities and 'extra charities' files combined, 212 have been identified not in fact to be charities (in most cases, the auto-matcher had been misled by the multiplicity of duplicate 'working names' Charities Commission records); a further seven had become charities during the course of the study, and one still remains unidentified.

A further 182 charities were identified among the records identified in the various attempts to 'cast the net wider' (plus 4 that became charities during the study period), and 19 more in the other minor additions to the file described here, for a total of 1826 suppliers that had been registered as charities for all or most of the period they'd been supplying the NHS.

Of those, only 44 in fact appear to be unincorporated: just over 2%, validating the original expectation that any charity supplying the NHS would seek the security of incorporation. Just under half the 44 that remain unincorporated are, in fact, NHS-associated charities: the charitable funds of NHS Trusts and hospitals themselves. It should be noted though that supplier charity incorporation is an ongoing process: over a dozen of the charities only registered as companies after making their first known transaction with the NHS, and several more not long before.

The legal status of organisations is also fluid: not only have previously unincorporated charities incorporated during the study period, more are adopting the specific status of CIO, and a handful have also changed status to/from CIC during the study period. All of these changes are recorded and detailed in the file.

This 'status fluidity' has also had an impact on the number of records in the file: sometimes the changes an organisation has undergone have resulted in it appearing (to the auto-matcher) as two (or more) different suppliers, only one of which had been included in the original source files for this work. In a handful of such cases it has been found beneficial to include the 'other' supplier(s) in the file too, and in such cases the extra record, since its inclusion was the direct consequence of a record already in one of the source files, has also been marked as originating from that source file.

But, as noted in Section 1.1, repair of corrupted data from a couple of the original NHS data sources has permitted removal from the file of records for now-redundant entangled supplier names; consequently the results file described here now includes only 1790 records identified as originating from the Charities source file, but 243 from the CIC source file. Every record in the results file is nonetheless identified by a unique and appropriate 'supplier' name by which it can be matched back to other original data.

The overall results of the matching process described in the preceding paragraphs can be summarised thus:

	Original			Results					
source file	charity only	charity & company	CIC	charity only	charity & company	CIC	mix	never CIC or charity	total
Charities	630	1087	0	22	1540	9	21	197+1unk	1790
CIC	0	0	241	0	0	232	1	10	243
extra Char	43		0	0	35	0	3	5	43
Manual	-	-	-	0	4	2	0	0	6
Manual 2	-	-	-	3	178	21	5	209	416
Man. CIC	-	-	-	0	7	29	0	7	43
CLWadd	-	-	-	1	7	1	0	1	10
total	-	-	241	26	1771	294	30	429+1unk	2551

'mix' comprises 6 organisations that were both charities and CICs at different times during the study period, 23 that are NHS-associated charities (5 of them incorporated for all or part of the time), and one that seems to have been an NHS GP surgery that then registered as a CIC.

4. Matching: special cases

4.1 GP surgeries

The source files included a number of suppliers which were in fact GP surgeries, for very few of which company registrations could be found (one maybe also as a charity). Services at many GP surgeries are supplied by a 'provider', and those providers (as listed for each surgery by CQC) fall primarily into two categories: they are either a distinct organisation, or the surgery itself. The handful of company-registered surgeries all fell into that latter category. In contrast, in the former case, a corporate registration was usually found, whether the provider serviced only a single surgery, or a local group, or a larger number over a wider area.

Those observations led to the conclusion that the 'public face' of an (NHS) GP surgery is never registered *per se*, although the provider may be. Accordingly, it is the provider for which a registration has been sought; generally one has been found for multi-site providers, as well as those few cases of 'self-providing' standalone surgeries (in at least one case by identifying a company of which the surgery's GPs were all directors). 'Self-providing' surgeries for which no registration can be found are presumed to operate under the umbrella of the NHS, and have been categorised as such.

4.2 Name and form changes

It is mentioned elsewhere that a number of the suppliers in the file were found to have changed their legal form during the study period, several of them indeed within the timespan of their NHS transactions. This section details how such changes are accommodated in the data file.

Those changes took various forms, and some involved a minor change of name, for example from Supplier Ltd to Supplier CIC. In many cases NHS customers continued recording transactions by the supplier's previous name long after the change had been registered.

But in a number of cases, some or all of the NHS customers eventually caught up with the change and started recording transactions by the supplier's new name. Even in the case of a single customer, this could in effect result in three 'phases' in the transactions records: continuing the above example, (i) Supplier Ltd and recorded as such; (ii) Supplier CIC but still recorded as Ltd; and (iii) eventually, Supplier CIC recorded as such.

It will be seen that (i) and (ii) will appear as the same supplier in the overall dataset, while (iii) will appear as a different supplier; but the aggregate number and value of transactions with each supplier will not correctly reflect the actual totals for the supplier as a whole, nor even for each legal form: in the present example, some of the transactions with the CIC will be missed. The situation of course becomes even more complex if the dates at which different NHS customers change the way they record the same supplier's name are also different.

As mentioned, such form changes affecting suppliers already present in the original 'checking' extracts are accommodated within the results file; and the work done to 'cast the net wider,' has found 'other' forms of several more suppliers, in cases when the name change had been relatively minor. However, cases have also been found in which the name has changed beyond the scope of patternwise automated recognition, and it was beyond the scope of the present work to seek all of those within the full dataset. The present example makes clear that it would be worthwhile to do so in certain cases, notably those which only adopted a Third Sector form during the study period.

5. File contents

The Stata file structure, as rendered by the 'describe' command, appears below, followed by a description of the purpose and content of each variable in turn. The file structure, variable definitions, and categorical definitions remain unchanged from the previous file, although of course counts and contents have changed.

```
Contains data from Charity and CIC checking Results v2.dta
  obs:         2,551
 vars:            30                          29 Jan 2021 21:59
 size:     1,535,702 (99.8% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
refno           long   %12.0g
supplier        str74  %74s
charityregno    long   %12.0g                 CharityRegNo
charitysubno    byte   %8.0g                  CharitySubNo
charregname     str80  %80s
charityname     str69  %69s                   CharityName
matchcodeCC     byte   %40.0g      matchcodeCC
CCnotes         str105 %105s
companynumber   str8   %9s                    CompanyNumber
companyname     str74  %74s                   CompanyName
matchcodeCH     byte   %44.0g      matchcodeCH
CHnotes         str120 %120s
orgtypes        byte   %25.0g      orgtypes
isCharity       byte   %16.0g      isCharity
isCompany       byte   %16.0g      isCompany
isCIC           byte   %12.0g      isCIC
isNHS           byte   %14.0g      isNHS
date_dep        byte   %9.0g       yesno
multi_supp      byte   %9.0g       yesno
count_supplier  int    %8.0g
amount_supplier double %12.0g
match_type      str22  %22s
matchsummCC     byte   %20.0g      matchsummCHCC
matchsummCH     byte   %20.0g      matchsummCHCC
matchsumm       byte   %22.0g      matchsumm
chkdSRB         byte   %11.0g      chkdSRB
sourcefile      byte   %17.0g      sourcefile
```

```
regno_old       float   %9.0g
subno_old       byte    %9.0g
compno_old      str8    %9s
-------------------------------------------------------------------------------
Sorted by:  supplier
```

Variable descriptions and contents

Original Identifiers

refno The unlabelled reference number associated with each supplier in the original files as supplied; for additional records, the equivalent number has been derived from the relevant transaction master list file (CCG, Trust or NHS England) and offset as appropriate, where possible. The three records without values have been included as a contingency: in two cases, a record has been created for the 'clean' version of a supplier name previously known only from one of the corrupted CCG datasets, while in the third, the organisation supplied the NHS when it was a charity, but has since become a CIC. Thus, if it is engaged in further transactions in its present form, the necessary matches are already available in the file. The ten previously unknown Third Sector suppliers added at the eleventh hour from the Central London (Westminster) CCG data are coded -1 in this variable, in the absence of any 'true' value for data not yet assimilated in the main database.

supplier The supplier name as it appears in the original files, and in the transaction lists and master auto-matching results file.

Details of matched Charity, if any

charityregno The registration number of the matched charity, if any. -1 is used when the organisation is known not to have been a charity during the study period. 'Missing' in the record for the one organisation that remains unidentified: it might yet turn out to be a charity (**matchcodeCC** 12).

charityregno	count
valid	1816
-1	734
[missing]	1

charitysubno The number of the matched sub-charity; most are zero. Takes the same value as **charityregno** for non-charities.

charitysubno	count
0	1806
1..9	10
-1	734
[missing]	1

charregname The registered name of the charity as shown by the Charities Commission online look-up at the time the charity was most recently checked. Several of these have changed in the course of the work, and more since the charity last transacted with the NHS. **CCnotes** will generally contain the explanation of any apparent mismatches.

charityname In the case of an un-revised auto-match, the charity name - whether registered, previous or 'working' - by which the auto-matcher identified the charity. In manual matches and rematches an exact match is unlikely of course (the auto-matcher would have found one if it existed) so this variable will hold a suitable version of the name of the matched charity.

matchcodeCC The main categorical variable indicating the results of the charity match-checking process. Its most important values are those which indicate a complex match, as detailed below the table. Values which indicate something about the nature of the matched organisations (codes 6 and up) take

precedence over lowered-numbered codes:

matchcodeCC	Freq.	Percent	Cum.
-9 no evidence of charity	264	10.35	10.35
-6 not charity, other	305	11.96	22.30
-2 not charity, imperf match removed	133	5.21	27.52
-1 not charity, perf match removed	21	0.82	28.34
1 existing perfect name match	1,144	44.85	73.19
2 existing perfect except 'LTD/LIMITED'	71	2.78	75.97
3 existing accepted imperfect name matc	186	7.29	83.26
4 corrected now perf name match	145	5.68	88.95
5 corrected imperf name match	187	7.33	96.28
6 match options	5	0.20	96.47
7 tentative match	7	0.27	96.75
8 date-dependent match	61	2.39	99.14
11 multi-supplier record	21	0.82	99.96
12 insufficient information	1	0.04	100.00
Total	2,551	100.00	

Detail of the categories:

-9 no evidence of charity — No evidence can be found to suggest that the supplier was a charity during the study period.

-6 not charity, other — The organisation has been identified as a type that is not a charity.

-2 not charity, imperf[ect] match removed — The organisation is not a charity, and an imperfect auto-match has been removed.

-1 not charity, perf[ect] match removed — The organisation is not a charity, and an auto-match has been removed, even though an apparently perfect name match. Either a charity closed before the NHS transactions in question, or an auto-match to a ‘working name’ that blocked a correct match elsewhere.

1 existing perfect name match — Auto-match was accepted which perfectly matched the supplier name, give or take ‘The’, &/and, &c

2 existing perfect except 'LTD/LIMITED' — Auto-match was accepted which perfectly matched the supplier name as in case 1, except for the omission/addition of LTD or LIMITED.

3 existing accepted imperfect name match — Auto-match was accepted which was not a perfect match for the given supplier name (other than cases 1 and 2 above).

4 corrected now perf[ect] name match — Used both when an auto-match has been revised, and when no prior auto-match existed. The definition of ‘perfect’ is equivalent to that used in cases 1 and 2 combined.

5 corrected imperf[ect] name match — Used both when an auto-match has been revised, and when no prior auto-match existed. Covers any new/revised match too ‘imperfect’ for case 4.

6 match options — It has been impossible to decide between two or more alternative matches. The one considered most likely is recorded in **charityregno-charityname**, the others detailed in **CCnotes**. Used whenever needed, regardless of whether any of the options was already an auto-match.

7 tentative match — The match recorded in **charityregno-charityname** seems likely, but cannot be proven with satisfactory confidence. Used whenever needed, regardless of any existing auto-match.

8 date-dependent match — The matched organisation changed its legal form or registration details during the period of the NHS transactions, or more recently (encoded thus in anticipation of further data acquisition). The match recorded in **charityregno-charityname** is the one current for the majority of NHS payments if possible; the other is detailed in **CCnotes**, along with changeover dates. Most commonly arises when a previously unincorporated charity incorporates and re-registers, though a number

	of other variations arise. Used whenever needed, regardless of any existing auto-match.
11 multi-supplier record	It is clear that two or more NHS customers used the same term to refer to different suppliers; perhaps each to their local branch of Mind, for example. The one with the most transactions is recorded in **charityregno-charityname**, and details of all the others are given in **CCnotes**. Used whenever needed, regardless of any existing auto-match.
12 insufficient information	The full weight of information available is still insufficient to identify the supplier. Used whenever needed, regardless of any existing auto-match.

(**matchcodeCC** and **matchcodeCH** use a unified coding scheme, and consequently not all codes appear in each table. Certain other 'working' codes have been also entirely superseded as the checking has progressed.)

CCnotes Free-text used to record any additional useful information about the match or matches recorded in **charityregno-matchcodeCC**. Terse and highly abbreviated when necessary; may be continued into the equivalent **CHnotes** at times. It is hoped that all the abbreviations will be clear with a little familiarity and comparison between records; explanations and clarifications upon request otherwise.

Details of matched Company, if any

[It should be noted that in a handful of cases of large multi-company conglomerates, it was very difficult to work out which was the best to record here. In cases where the 'right' company could not be clearly identified - or if the 'right'-looking one was not financially active to a sufficient degree - the supplier is matched to the group parent company.]

companynumber The registration number of the matched company, if any. If details of a charity are recorded in **charityregno-charityname**, then the registration recorded here should *always* correspond to that charity, regardless of any other consideration. Note that the numbers for certain types include non-numeric prefices: CE (for CIO) and IP (for IPS) are most common here. "-1" is used when the organisation is known not to have been a company during the study period. 'Missing' in the 17 cases when it has not been possible to identify the supplier, and in the five in which the supplier organisation is identified but its incorporated status (and thus registration) cannot be found (**matchcodeCH**s 12 and 13 respectively). Four further companies have been clearly identified, but appear not to be registered in the UK: those records also bear **matchcodeCH** 13, but are specifically identified by "NoneInUK" in this (**companynumber**) variable.

companynumber	count
valid	2350
-1	174
[blank]	23
NoneInUK	4

companyname The registered name of the company as shown by the Companies House online look-up at the time the company was most recently checked. Several of these have changed in the course of the checking, and more since the company last transacted with the NHS. **CHnotes** will generally contain the explanation of any apparent mismatches. If details of a charity are recorded in **charityregno-charityname**, then the registered company name recorded here should *always* correspond to that charity, regardless of any other consideration.

matchcodeCH The main categorical variable indicating the results of the company match-checking process. Its most important values are those which indicate a complex match, as detailed below the table. Values which indicate something about the nature of the matched organisations (codes 6 and up) take precedence over lowered-numbered codes:

matchcodeCH	Freq.	Percent	Cum.
-9 no evidence of company	25	0.98	0.98
-6 not company, other	25	0.98	1.96
-5 not company, NHS-assoc	27	1.06	3.02
-4 not company, NHS	82	3.21	6.23
1 existing perfect name match	1,083	42.45	48.69
2 existing perfect except 'LTD/CIC/LIMI	112	4.39	53.08
3 existing accepted imperfect name matc	163	6.39	59.47
4 corrected now perf name match	277	10.86	70.33
5 corrected imperf name match	595	23.32	93.65
6 match options	6	0.24	93.88
7 tentative match	30	1.18	95.06
8 date-dependent match	75	2.94	98.00
11 multi-supplier record	22	0.86	98.86
12 insufficient information	19	0.74	99.61
13 identified but not found	10	0.39	100.00
Total	2,551	100.00	

Detail of the categories:

-9 no evidence of company — No evidence can be found to suggest that the supplier was a registered company during the study period.

-6 not company, other — The organisation has been identified as a type (other than an NHS-associated type) that is not a company.

-5 not company, NHS-assoc[iated] — The organisation is not a company, but has been identified as being NHS-associated; primarily the Charitable Funds of NHS institutions, which are typically registered as charities but not as companies. Excludes organisations in case -4.

-4 not company, NHS — The organisation is an NHS institution, such as a PCT or other Trust, or a specialised unit within a hospital, or sometimes a GP surgery that appears to be registered in no other way.

1 existing perfect name match — Auto-match was accepted which perfectly matched the supplier name, give or take ‘The’, &/and, &c

2 existing perfect except 'LTD/CIC/LIMI[TED'] — Auto-match was accepted which perfectly matched the supplier name as in case 1, except for the omission/addition of CIC, LTD, or LIMITED.

3 existing accepted imperfect name match — Auto-match was accepted which was not a perfect match for the given supplier name (other than cases 1 and 2 above).

4 corrected now perf[ect] name match — Used both when an auto-match has been revised, and when no prior auto-match existed. The definition of ‘perfect’ is equivalent to that used in cases 1 and 2 combined.

5 corrected imperf[ect] name match — Used both when an auto-match has been revised, and when no prior auto-match existed. Covers any new/revised match too ‘imperfect’ for case 4.

6 match options — It has been impossible to decide between two or more alternative matches. The one considered most likely is recorded in **companynumber** and **companyname**, the others detailed in **CHnotes**. Used whenever needed, regardless of whether any of the options was already an auto-match.

7 tentative match — The match recorded in **companynumber** and **companyname** seems likely, but cannot be proven with satisfactory confidence. Used whenever needed, regardless of any existing auto-match.

8 date-dependent match — The matched organisation changed its legal form or registration details during the period of the NHS transactions, or more recently (encoded thus in anticipation of further data acquisition). The match recorded in **companynumber** and **companyname** is the one current for the majority of NHS payments if possible; the other is detailed in **CHnotes**, along with changeover dates. Most commonly arises when a

previously unincorporated charity incorporates and re-registers; if in such a case most NHS transactions were made before incorporation, then **companynumber** and **companyname** will indicate no company match (-1 and blank respectively), because the charity was not registered as a company at the relevant time; details of the eventual registration (often as a CIO) are in **CHnotes**. Other variations also arise though. Used whenever needed, regardless of any existing auto-match.

11 multi-supplier record — It is clear that two or more NHS customers used the same term to refer to different organisations; perhaps each to their local branch of Mind, for example. The one with the most transactions is recorded in **companynumber** and **companyname**, and details of all the others are given in **CHnotes**. Used whenever needed, regardless of any existing auto-match.

12 insufficient information — The full weight of information available is still insufficient to identify the supplier. Used whenever needed, regardless of any existing auto-match.

13 identified but not found — Even though the supplier has been identified as an organisation, nonetheless a company registration cannot be found. Also used in the four cases of companies which appear only to be registered overseas.

(**matchcodeCC** and **matchcodeCH** use a unified coding scheme, and consequently not all codes appear in each table. Certain other 'working' codes have been also entirely superseded as the checking has progressed.)

CHnotes Free-text used to record any additional useful information about the match or matches recorded in **companynumber**-**matchcodeCH**. Terse and highly abbreviated when necessary; may be continued from the equivalent **CCnotes** at times. It is hoped that all the abbreviations will be clear with a little familiarity and comparison between records; explanations and clarifications upon request otherwise.

Characteristics of the suppliers

orgtypes Categorical variable to indicate all the basic legal forms which the supplier is known to have taken during the study period; not necessarily simultaneously:

orgtypes	Freq.	Percent	Cum.
0 none ID'd	43	1.69	1.69
1 charity	26	1.02	2.70
2 company (not CIC)	295	11.56	14.27
3 charity, company	1,771	69.42	83.69
4 CIC	294	11.52	95.22
5 CIC, charity	6	0.24	95.45
8 NHS or assoc	91	3.57	99.02
9 charity, NHS	18	0.71	99.73
10 comp, NHS	1	0.04	99.76
11 charity, comp, NHS	5	0.20	99.96
12 CIC, NHS	1	0.04	100.00
Total	2,551	100.00	

Notes on selected table entries:

2 company (not CIC) — Specifically non-charitable, non-Community Interest, companies. Ordinary incorporated commercial concerns, in other words.

3 charity, company — Primarily, charitable suppliers that were also company-registered for all, or perhaps only the latter part, of the study period.

5 CIC, charity — Several organisations were found that had swapped form between charity and CIC. And although, in theory, a charity cannot register as a CIC, one of these suppliers (X-Pert Health) seems somehow to have had parts of itself registered as both simultaneously. The CIC part

	appears to be the better match.
9 charity, NHS	Usually the Charitable Funds of NHS Trusts.
10 comp, NHS	The sole case is a GP surgery. It appears to have functioned as an NHS entity (shown by CQC as its own 'provider') for which no company registration could be found, until late in the study period when it registered as a private limited company.
11 charity, comp, NHS	Two of these records refer to Imperial College Healthcare Charity, which seems to have started as a charity-registered NHS operation and then later been incorporated, during the study period. The other three are NHS Hospital support charities which, unusually, are also registered companies.
12 CIC, NHS	The sole case is a GP surgery. It appears to have functioned as an NHS entity (shown by CQC as its own 'provider') for which no company registration could be found, until late in the study period when it registered as a CIC.

(The coding is binary by type, so that the code for each multi-type category is the sum of the codes of its component types: 5 = 1+4 for example, or 11 = 1+2+8 = 3+8 &c.)

isCharity In concept, a flag to indicate whether this supplier was ever a charity during the time it was transacting with the NHS. But there are 11 more suppliers which have since registered as charities, and thus might join the list as charitable suppliers if further transaction data are gathered. There's also one other the charitable status of which cannot be determined: it could be one of several organisations, charities among them.

isCharity	Freq.	Percent	Cum.
-1 unknown	1	0.04	0.04
0 not charity	724	28.38	28.42
1 charity	1,815	71.15	99.57
2 became charity	11	0.43	100.00
Total	2,551	100.00	

isCompany Like **isCharity**, conceptually a flag to indicate whether a supplier was ever a registered company during the time it was transacting with the NHS. But there are 17 more suppliers which have since incorporated, and thus might join the list as such if further transaction data are gathered. There's also one other the status of which cannot be determined: it could be one of several organisations, not all of which seem to be incorporated.

isCompany	Freq.	Percent	Cum.
-1 unknown	20	0.78	0.78
0 not company	158	6.19	6.98
1 company	2,356	92.36	99.33
2 became company	17	0.67	100.00
Total	2,551	100.00	

isCIC Like the two preceding variables, conceptually a flag to indicate whether a supplier was ever a Community Interest Company during the time it was transacting with the NHS. A further 6 suppliers have since registered that form, and thus might join the list as such if further transaction data are gathered.

isCIC	Freq.	Percent	Cum.
0 not CIC	2,250	88.20	88.20
1 CIC	295	11.56	99.76
2 became CIC	6	0.24	100.00
Total	2,551	100.00	

isNHS Like the three preceding variables, conceptually a flag to indicate whether a supplier was ever an NHS entity during the period of its transactions. There are also four others the status of which cannot be determined, but which could be internal NHS entities or processes.

isNHS	Freq.	Percent	Cum.
-1 unknown	4	0.16	0.16
0 not NHS	2,431	95.30	95.45
1 NHS or assoc	116	4.55	100.00
Total	2,551	100.00	

date_dep A simple flag to identify those suppliers whose registration details changed during or immediately after the period of their NHS transactions (ie, those coded 8 in **matchcodeCC** or **matchcodeCH**).

date_dep	Freq.	Percent	Cum.
0 no	2,463	96.55	96.55
1 yes	88	3.45	100.00
Total	2,551	100.00	

multi_supp A simple flag to identify those records which actually represent several different suppliers, each known to its local NHS customers by the same name (ie, those coded 11 in **matchcodeCC** or **matchcodeCH**).

multi_supp	Freq.	Percent	Cum.
0 no	2,529	99.14	99.14
1 yes	22	0.86	100.00
Total	2,551	100.00	

count_supplier The count of the number of known NHS transactions with this supplier, included in the original files as supplied; for additional records, the equivalent count has been totalled from the relevant transaction master list file(s) (CCG, Trust and/or NHS England) where possible. The sole record with a zero has been included as a contingency: the organisation supplied the NHS when it was a charity, but has since become a CIC. If it is engaged in further transactions in its present form, the necessary matches are already available in the file. The twelve missing values comprise the two records for the 'true' supplier names previously known only from now-redundant corrupted records, and the ten last-minute additions from the recently-parsed Central London (Westminster) CCG data.

count_supplier	count
valid	2538
0	1
[missing]	12

amount_supplier The sum of the known NHS payments to this supplier, included in the original files as supplied; for additional records, the equivalent sum has been totalled from the relevant transaction master list file(s) (CCG, Trust and/or NHS England) where possible. Several records have zero values, including the 'contingency' one mentioned under **count_supplier** above. The twelve missing values comprise the two records for the 'true' supplier names previously known only from now-redundant corrupted records, and the ten last-minute additions from the recently-parsed Central London (Westminster) CCG data.

amount_supplier	count	
valid	2538	(includes zeroes)
[missing]	12	

Matching metadata

match_type A recoded and augmented version of the variable of the same name in the original files as supplied. In those, it indicated the source(s) from which any existing auto-match was derived. It has now been extended to record how those matches were augmented and revised, and contains many combinations of the following codes, colon-separated as in the original. The original entries have been abbreviated to permit inclusion of the new ones:

Original entries

present code	count	original form	interpretation
CC	1790	Charity Commission	auto-matched from Charities Commission master lists
CH	1473	Companies House	auto-matched from Companies House master lists
NHSD	191	NHS Digital	auto-matched from NHS Digital master lists
Anm	238	No Match	no auto-match
NDoc	1	Named Doctor	auto-matched as an individual doctor (in fact, it's Dr Kershaw's Hospice)

Additional entries

code	count	expansion	interpretation
xCC	43	Extra Charities	a record from the 'extra charities' file
Mm	724	Manual match	match made where none had previously existed
Mnm	271	Manual no-match	existing auto-match removed and not replaced
Mrm	198	Manual rematch	existing auto-match replaced
Mdk	21	Manual don't-know	manual searches did not yield a matchable organisation
CH via CC	247		company match obtained from Charities Commission online look-up.

No code will appear in match_type more than once, even if, for example, pre-existing charity and company matches were both removed or replaced. The entries here are included perhaps more for interest and completeness than analysis, and tend not to tell the full story of some of the most challenging cases, in which matches have been made, remade, deleted, restored and augmented several times over.

matchsummCC Summarises the status of the supplier's Charities Commission matching.

matchsummCC	Freq.	Percent	Cum.
0 not matchable	546	21.40	21.40
1 accepted automatch	1,403	55.00	76.40
2 revised match	102	4.00	80.40
3 enhanced match	322	12.62	93.02
4 removed automatch	178	6.98	100.00
Total	2,551	100.00	

case 3 applies when a match was found for a supplier without an auto-match, and also in cases of multiple matches: **matchcodeCC** cases '6 options', '8 date-dependent', and '11 multi-supplier'.

matchsummCH Summarises the status of the supplier's Companies House matching.

matchsummCH	Freq.	Percent	Cum.
0 not matchable	179	7.02	7.02
1 accepted automatch	1,358	53.23	60.25
2 revised match	66	2.59	62.84
3 enhanced match	939	36.81	99.65
4 removed automatch	9	0.35	100.00
Total	2,551	100.00	

case 3 applies when a match was found for a supplier without an auto-match, and also in cases of multiple matches: **matchcodeCH** cases '6 options', '8 date-dependent', and '11 multi-supplier'.

matchsumm Summarises the status of the supplier's overall matching. Derived from the preceding **matchsummCC** and **matchsummCH**.

matchsumm	Freq.	Percent	Cum.
0 not matchable	66	2.59	2.59
1 accepted automatch	1,345	52.72	55.31
2 revised match	57	2.23	57.55
3 enhanced match	1,006	39.44	96.98
4 removed automatch/es	77	3.02	100.00
Total	2,551	100.00	

cases:
- 0 unmatchable in both CC and CH
- 1 CC and CH auto-match both accepted unchanged (even if one of them was a no-match)
- 2 auto-match revised in CC, CH or both (includes removal of one and revision of the other)
- 3 additional match in CC, CH or both (ie, case 3 in either component variable)
- 4 CC and/or CH auto-matches removed to leave no-match in both.

chkdSRB Degree of manual checking of the match itself. All matches have of course been assessed in multiple different ways for need of checking, and all additions to the file beyond the original Charities and CICs pair have in any case been fully manually matched.

chkdSRB	Freq.	Percent	Cum.
0 not chkd	256	10.04	10.04
1 manual	1,901	74.52	84.56
2 cursory	255	10.00	94.55
3 curs+geog	139	5.45	100.00
Total	2,551	100.00	

cases:
- 0 Not manually checked. These suppliers have perfect, unambiguous and usually geographically-specific name matches, and have passed automatic comparisons of their **amount_supplier** with Charities Commission income data. Additionally, their matches have registration numbers low enough to indicate that they were registered no later than the start of the study period. (Both CC and CH allocate registration numbers sequentially: the higher the number, the later the registration.)
- 1 Full manual check. The supplier has been identified and matched through online research based on name, customer locations, expense type and area entries, and anything else that could be gleaned from the source files. The resulting match is then compared with any existing auto-match, and any discrepancy further investigated until it is certain which is right, if it can be.
- 2 The existing match has been examined for viability, and accepted, or lightly revised (typically, for example, if recently registered). These are the cases in which the check started with the existing match. Any uncertainty arising led, of course, to a full 'case 1' investigation: the only way to rule out an existing match is to find a better one.
- 3 In addition to the checks in case 2, a detailed comparison has been made between the location of the matched organisation and the location(s) of its NHS customer(s). Typically sufficient when the umbrella organisation is unmistakable (eg Mind, Age UK, Home-Start) but the appropriate branch may not have been found. Often, in cases of short and cryptic acronyms, the starting point for a full 'case 1' investigation when a geographical discrepancy was found.

Data origins

sourcefile A categorical variable indicating the source of the record. As discussed elsewhere, a handful of entries in the original Charities and CICs files were best dealt with by introducing an additional record for a name variant of the same organisation, but subsequently it has been possible to remove certain 'entangled' supplier names from the database entirely; the net effect of those has been a reduction in the number of records derived from the original Charities file, but an increase for CICs (initially 1806 and 241 records respectively).

sourcefile	Freq.	Percent	Cum.
1 Charities	1,790	70.17	70.17
2 CICs	243	9.53	79.69
3 Extra Charities	43	1.69	81.38
4 Manual Addition	6	0.24	81.62
5 Manual Addn 2	416	16.31	97.92
6 Manual CIC Addn	43	1.69	99.61
7 CLW CCG Addn	10	0.39	100.00
Total	2,551	100.00	

The terse later codes my be interpreted as follows:

5 Manual Addn 2	The records identified by the searches in the wider database for Third Sector suppliers that might have been missed by the auto-matching process, as detailed in Section 1.1.
6 Manual CIC Addn	The further records identified by the searches in the wider database specifically for suppliers with names similar enough to those of known CICs to warrant further investigation, as detailed in Section 1.1.
7 CLW CCG Addn	The previously-unknown suppliers within the last-minute Central London (Westminster) data that were found to be Third Sector.

regno_old The Charities Commission registration number of any auto-match in the source file as received.

subno_old The sub-charity number, if appropriate, of any charity auto-match in the source file as received.

compno_old The Companies House registration number of any auto-match in the source file as received.

Omissions

The astute reader will have observed that certain variables from the source files have not been included in the results file, in particular charitynameno, and the count_verif, amount_verif pair. Taking them in turn, the first, charitynameno, has been omitted for a couple of reasons. Firstly, it is of extremely limited utility in further analysis owing to high level of name duplication within the CC 'Names' extract, and even within the entries for the same charity; and secondly it is not used or reported within CC's online portal which was a major resource for checking and correction of matches within this work. That portal will list previous and 'working' names, but doesn't number them, nor report duplicates even when they exist in the database. Consequently, the 'nameno' for any revised or augmented match was not apparent, and although it could be matched back into the file from the look-up, the benefit of doing so was not apparent, and certainly not to an extent worth the cost in time and extra space in the file: as it is the presence of **charregname** as well as **charityname** would require two 'nameno' entries, while any date-dependent or multi-supplier match might require at least two more. Finally, out of context a nameno cannot easily be distinguished from a Regno or Company Number and consequently opens a route to mismatches. In contrast, the distinction between Regno (always 6 or 7 digits) and Company Number (always 8 characters, including leading zeroes or alphabetical prefices) is clear.

The count_verif, amount_verif pair have been omitted because it is not possible to derive correct values for them for all cases in the present file. Values could be generated, but they would not be right in either

the date-dependent (**matchcode** 8) or multi-supplier (**matchcode** 11) records, where too many transactions would be associated with the 'primary' supplier, nor in cases of a change in legal form like those described in Section 4.2, in which either too many, or too few, transactions might be included depending on the date and type of change, and which 'versions' of the supplier organisation were present and absent in the file. Further, the necessary data are not yet available from the recovered Surrey Downs CCG files, nor the recently-extracted Central London (Westminster) CCG ones.

6. Further Work

It is believed that relatively few further Third Sector supplies now remain undetected within the wider dataset; the various search algorithms described in Section 1.1 reached the point where they were cycling over and over the same results, and it's now some while since a 'new' Third Sector supplier (or candidate) has been spotted by chance within the wider dataset. Nonetheless, a few will inevitably remain, and a diligent search of those suppliers with the largest transaction totals but imperfect or non-existent auto-matches might yield the most significant among them.

As the Central London (Westminster) CCG data also revealed when they became available at the last minute, any new data source is likely to bring with it new 'local' Third Sector suppliers. While the file here described includes a number of large national charities, the majority of the Third Sector suppliers therein are local - often very local - to their NHS customers, so any further customers added to the database are likely to bring with them further Third Sector suppliers.