BMJ Open

Linking population-based cohorts with cancer registries in LMIC: a case study and lessons learnt in India

Aastha Aggarwal ,[1,2] Ranganathan Rama,[3] Preet K Dhillon,[1,2,4] Mohan Deepa,[5] Dimple Kondal,[2] Naveen Kaushik,[2] Dipika Bumb,[6] Ravi Mehrotra,[7,8] Betsy A Kohler,[9] Viswanathan Mohan,[5,10] Theresa W Gillespie ,[11,12] Alpa V Patel,[13] Swaminathan Rajaraman,[3] Dorairaj Prabhakaran,[1,2] Kevin C Ward,[8,12,14] Michael Goodman [8,12,14]

To cite: Aggarwal A, Rama R, Dhillon PK, *et al*. Linking population-based cohorts with cancer registries in LMIC: a case study and lessons learnt in India. *BMJ Open* 2023;**13**:e068644. doi:10.1136/bmjopen-2022-068644

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-068644).

Received 27 September 2022
Accepted 15 February 2023

For numbered affiliations see end of article.

Correspondence to
Dr Aastha Aggarwal;
aastha.aggarwal@phfi.org

ABSTRACT

Objectives In resource-constrained settings, cancer epidemiology research typically relies on self-reported diagnoses. To test a more systematic alternative approach, we assessed the feasibility of linking a cohort with a cancer registry.

Setting Data linkage was performed between a population-based cohort in Chennai, India, with a local population-based cancer registry.

Participants Data set of Centre for Cardiometabolic Risk Reduction in South-Asia (CARRS) cohort participants (N=11 772) from Chennai was linked with the cancer registry data set for the period 1982–2015 (N=140 986).

Methods and outcome measures Match*Pro, a probabilistic record linkage software, was used for computerised linkages followed by manual review of high scoring records. The variables used for linkage included participant name, gender, age, address, Postal Index Number and father's and spouse's name. Registry records between 2010 and 2015 and between 1982 and 2015, respectively, represented incident and all (both incident and prevalent) cases. The extent of agreement between self-reports and registry-based ascertainment was expressed as the proportion of cases found in both data sets among cases identified independently in each source.

Results There were 52 self-reported cancer cases among 11 772 cohort participants, but 5 cases were misreported. Of the remaining 47 eligible self-reported cases (incident and prevalent), 37 (79%) were confirmed by registry linkage. Among 29 self-reported incident cancers, 25 (86%) were found in the registry. Registry linkage also identified 24 previously not reported cancers; 12 of those were incident cases. The likelihood of linkage was higher in more recent years (2014–2015).

Conclusions Although linkage variables in this study had limited discriminatory power in the absence of a unique identifier, an appreciable proportion of self-reported cases were confirmed in the registry via linkages. More importantly, the linkages also identified many previously unreported cases. These findings offer new insights that can inform future cancer surveillance and research in low-income and middle-income countries.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This study linked data from two structured and well-maintained data sources—the population-based Chennai cancer registry, and population-based cohort, CARRS (Centre for Cardiometabolic Risk Reduction in South-Asia).
⇒ This study made use of an open access probabilistic record linkage software originally developed for use in the data from western populations.
⇒ A linkage protocol was developed using linkage variables with limited discriminatory power and adjusted to nuances of data collected in India, a low-income to middle-income country, which can be modified for use in similar settings.
⇒ A unique identifier such as Social Security Number was not available for use as a linkage variable which could have further strengthened the linkage protocol.
⇒ Missed-matches due to the selection of a threshold score presented a potential source of bias. However, even a marginally lower threshold (eg, from 20 to 15) would have rendered this study practically impossible as it would have nearly tripled the number of records requiring manual review. Confirmation of unlinked self-reported cases was dependent only on collection of additional information from the participants as lack of access rendered linkages with vital registration system unfeasible.

INTRODUCTION

An important methodological consideration in prospective studies of cancer incidence is the accurate ascertainment of new malignancies diagnosed during follow-up. While self-reports are routinely used to identify incident cancers, they are subject to error.[1–5] In high-income countries (HIC), linking study data with population-based cancer registries (PBCR) is a proven method of ascertaining incident cases and capturing clinical cancer characteristics.[6–11]

Cancer registries play an increasingly important role in low-income to middle-income countries (LMIC), especially in India, where the numbers of incident cases and cancer deaths have doubled from 1990 to 2016.[12] Under the National Cancer Registry Programme (NCRP) initiated by the Indian Council of Medical Research in 1981, there are currently 36 PBCRs and 236 hospital-based cancer registries.[13]

The availability of PBCR data in India offers a range of opportunities for epidemiological research, especially considering the increasing number of population-based studies conducted in different parts of the country.[14] While many of these studies are focused on health outcomes other than cancer, they could be leveraged to evaluate cancer incidence, cancer-specific mortality and difficult to obtain cancer clinical data among study participants if linked to PBCR data.

With this background in mind, the current study aimed to pilot test linkage between an ongoing population-based cohort study in Chennai with the large and well-established PBCR in that city, and also use information on self-reported cancer diagnoses in the cohort. The three specific research questions addressed by this study are as follows: (1) Can existing data linkage software packages developed in HIC be used in LMIC, such as India? (2) What proportion of self-reported incident cancer cases are found in the PBCR? and (3) How many cases that were not self-reported can be identified in the PBCR? The following sections address each of the above questions and provide recommendations for optimising data linkages in India and other LMIC.

MATERIALS AND METHODS

Study cohort

The study population represents one site—Chennai—of the Centre for Cardiometabolic Risk Reduction in South-Asia (CARRS), a population-based cohort established in three large urban centres of South Asia to measure prevalence and trends in cardiometabolic diseases through interviewer-administered questionnaires conducted mostly on an annual basis.[15] The study is composed of two separate, independently sampled cohorts: CARRS-1, established in 2010–2011, and CARRS-2, established in 2014–2015.

Self-reports of 'ever diagnosis' of cancer, year of diagnosis and the site of cancer were obtained as part of the study during the fourth and fifth follow-ups of CARRS-1 (2016–2018) and the baseline of CARRS-2 (2014–2016). Additional cancer cases were identified through verbal autopsy reports collected from next-of-kin. Participants provided informed written or verbal consent for linking their data with the disease registries. Verbal consent was taken from next-of-kin for verbal autopsy data. Participation in the cohort was not dependent on consent to link data with disease registries.

For this linkage, we only considered self-reported cases with date of diagnosis up to 2015 as the data in the cancer registry was available through the end of that year.

Cancer registry

Established in 1981, the Madras Metropolitan Tumour Registry (MMTR) covers an area of 170 km^2, comprising the city of Chennai, the largest metropolitan centre in the state of Tamil Nadu.[16] Per the 2011 Census, the population of Chennai city was 4 646 732, which constitutes 6.5% of the state of Tamil Nadu and 0.4% of the total population of India.[17 18] The registry primarily uses active surveillance methods and collects information from >250 clinical sites using standardised NCRP-PBCR forms.[16] In 2012–2016, a total of 31 271 cases were registered in the MMTR.[13]

For this linkage, we used available MMTR data from 1982 through 2015.

Linkage protocol

The primary linkage task was to match all CARRS participants against the MMTR data for 2010–2015. This was done to (1) validate incident cancer cases captured via self-report or verbal autopsy (hereafter collectively referred to as 'self-reported cases') and (2) ascertain additional incident cancer cases previously not reported among CARRS participants. The secondary linkage task was to perform a similar matching for all (both prevalent and incident) cases for the period 1982–2015. Incident cancers were defined as cases diagnosed during CARRS follow-up since enrolment in the study (figure 1).

Although the Unique National ID system (Aadhaar) is currently being implemented in India, the coverage is still incomplete. For this reason, the linkage protocol used a combination of variables with limited discriminatory power, including participant name, gender, age, address, Postal Index Number (PIN) code and father's and spouse's name. A person's current age was derived in both data sets using the year of birth in CARRS, and age and year of cancer diagnosis in the MMTR. Naming conventions in India are quite different from those used in most Western countries. Indian names are based on a variety of systems with strong influences from a geographical region, religion and caste. They usually constitute a given name and a variable number of secondary names. The secondary name could be a surname or a patronym, it may reflect the person's caste, occupation or place of origin, which is common in Tamil naming convention. Often, the place of origin and the patronym are initialised, but not spelt out.[19 20] In addition to these various factors, the order of these names is also variable. Gender was treated as a binary variable with 'Male' and 'Female' categories. PIN code is a six-digit code in the Indian postal code system. For some records, only the last two digits were recorded.

Linkage was carried out using Match*Pro software, a probabilistic data linkage programme based on the Fellegi and Sunter model, developed by Information Management Services and provided by the National Cancer Institute.[21] Match*Pro offers considerable flexibility in specifying blocking and matching parameters, adjusting weights, setting predefined scenarios for acceptable and

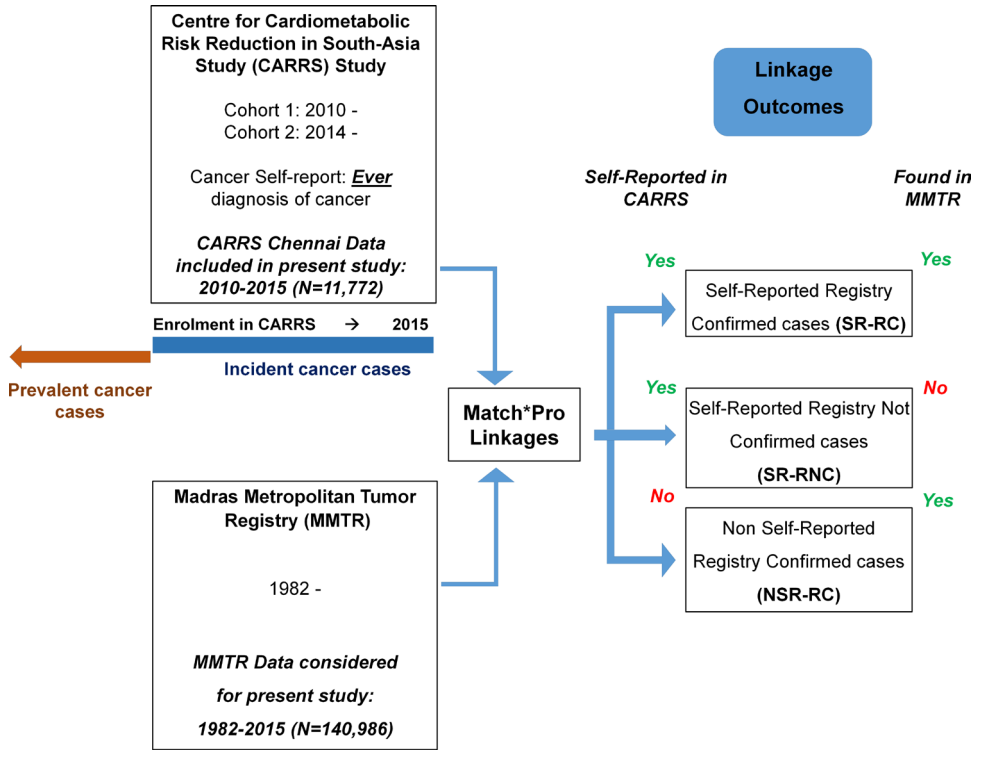

Figure 1 Details of study data sets and categories of linkage outcomes.

unacceptable matches and facilitates a manual review of output records. The use of blocking parameters increases linkage efficiency by limiting the number of comparisons to records where one or more parameters agree. The software assigns probability scores based on the agreement between records for matching parameters. A linkage configuration file consisting of blocking parameters (participant name, gender, father's name, spouse's name and PIN code) and matching parameters (participant name, gender, father/spouse name, current age, address and PIN code) along with the matching algorithm were iteratively optimised based on the local naming patterns and address structures (details in online supplemental file).

Names were recorded inconsistently in the two data sets: sometimes as a combination of initials and first name, and at other times full names were available. In MMTR, the father's and spouse's names were recorded as separate variables. However, in one round of data collection in CARRS, participants had the choice to provide their father's name or spouse's name under the same field, which made these two variables indistinguishable in CARRS-1 data. For this reason, these two variables were grouped together in the linkage configuration file.

Residential address was recorded in both the data sets. However, these were recorded in inconsistent patterns. Further, street and locality names were variably abbreviated. For some participants in CARRS, multiple addresses were available that were all used in the linkage configuration file.

One of the features of probabilistic linkages is the ability to account for the fact that no data is free of errors and inconsistencies. While Match*Pro's algorithm facilitates the handling of such data, preliminary cleaning of CARRS data was done to reduce missingness and errors in the variables of interest. Prefixes such as 'late', 'Mr', 'Shri', 'Shrimati', etc were removed. First name, middle name and last name were concatenated to generate a single name variable. Before doing this, efforts were made to deduce missing last names or expand the initial in the last name based on information on names of other household members. Sometimes the father's or spouse's full name was mentioned under the participant's last name. This was corrected. Also, if the last name was included both as an initial in the first name and in expanded form under last name the initial was removed from first name. For some of the participants, PIN code was recorded along with address variables. In such cases it was separated from the address variable as the PIN code was used as a separate variable for linkages. Whenever PIN code was recorded as a double-digit number, it was changed to the standard 6-digit format. If the exact date of birth was not available, year of birth was back calculated in CARRS data from the age at the time of the interview. Year of birth was also recorded in the MMTR database. These were then used to compute approximate age at the time of data linkage. The linkage process generated a score for potential matches, with higher scores indicating a higher likelihood of a true match. Following electronic linkage, the study team manually reviewed all matches with scores ≥20. As the assessment of linkages between CARRS and MMTR was contingent on validation of self-reports, optimisation of threshold minimum score for review was done by first investigating only self-reported cases by the

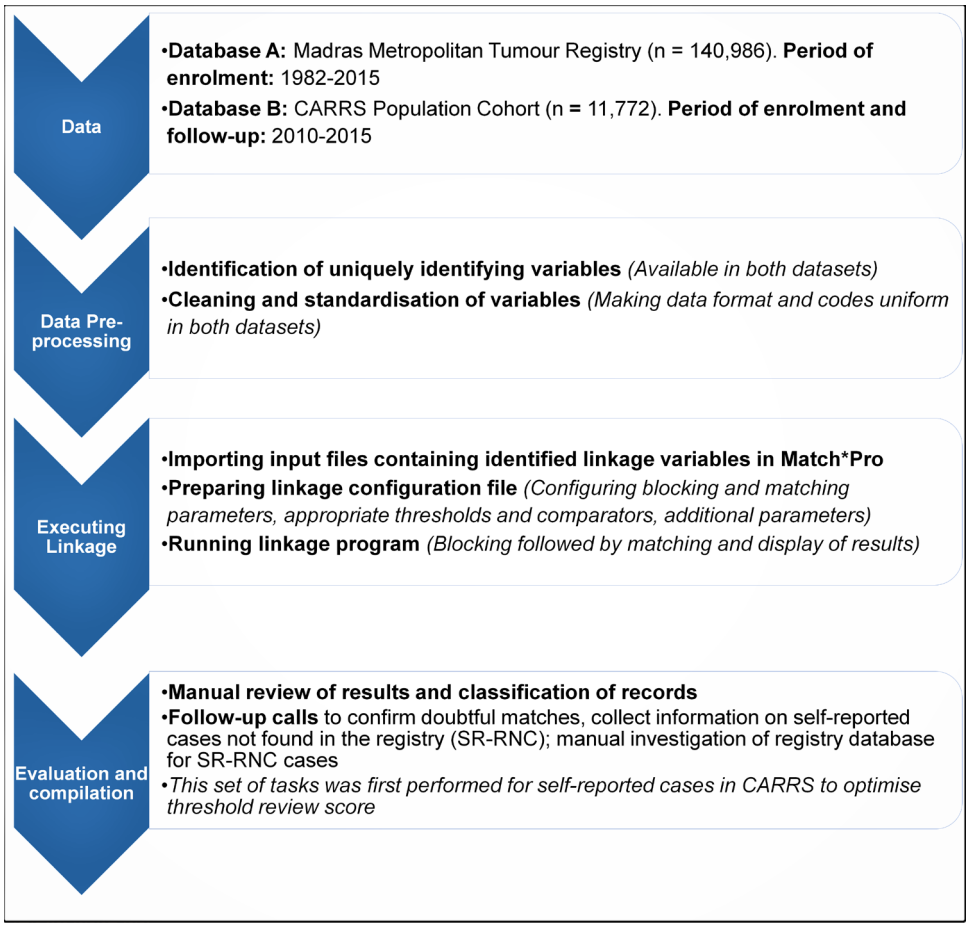

Figure 2 Description of the linkage process.

linkage process. In the output file with a score range of 12.8–56.1, we observed that 84% of self-reported cases linked to the registry had a score ≥20.

When available in both data sets, additional variables such as reported cancer site, year of diagnosis and relative's name were also used for manual review. Whenever possible, follow-up calls and additional reviews of medical records were used to collect information on cases not confirmed in the registry, and/or adjudicate doubtful matches for both self-reported and non-self-reported cases. When it was not possible to contact the participants or their next-of-kin by telephone, attempts were made to visit them. Self-reported cases not found via linkages were also manually investigated in the registry database. Steps involved in the linkage process are shown in figure 2.

Data files used for linkages included only the required variables. They were uploaded by registry personnel on a password-protected laptop designated for this purpose, and linkage was performed at the premises of the MMTR. No information could be digitally copied from this laptop. Information on potential matches was manually recorded for further action.

Results obtained from linkage were classified as follows: (1) Self-reported registry-confirmed (SR-RC) cases, that is, self-reported cases successfully confirmed in the registry via linkage; (2) self-reported cases not confirmed in the registry (SR-RNC), that is, self-reported cases not found in the registry via linkages; and (3) non-self-reported registry confirmed (NSR-RC) cases, that is, cases not self-reported by CARRS participants but identified in the registry via linkage (figure 1).

The CARRS and MMTR data sets were compared for data completeness for the main linkage variables. Treating the registry data as gold-standard, the agreement between self-reported and registry-based cancer case ascertainment was examined by estimating sensitivity, specificity and positive and negative predictive values (PPV and NPV), along with the kappa statistic; each measure accompanied by a 95% CI. We also compared the distributions of sociodemographic variables and wealth index among all CARRS participants and those cohort members who had a registry-confirmed (SR-RC and NSR-RC) incident cancer diagnosis. The wealth index was characterised by the availability of household amenities and assets.[22] The differences between SR-RC and NSR-RC incident cases with respect to the extent of agreement between CARRS and MMTR demographic variables were assessed using a two-proportion Z test. These two groups were also compared for the distribution of disease characteristics using Fisher's exact test. The analysis was performed using Stata V.15.

Patient and public involvement

It was not appropriate or possible to involve patients or the public in the design, or conduct, or reporting, or dissemination plans of our research.

RESULTS

The CARRS cohort data set included 11 772 individuals. The MMTR data set included 140 986 patients with cancer with diagnosis dates from 1982 through 2015. Of those, 35 763 patients with cancer had diagnosis dates from 2010 through 2015 and were considered incident cases. For the linkage variables, information was available in at least 99% of records in both data sets for a person's name, current age, gender, address and PIN code (data not shown). The data completeness differed between the two data sets for father's and spouse's name (>90% in CARRS and <45% for MMTR) among all subjects, and for the year of diagnosis (53% in CARRS and 100% for MMTR) among cancer cases.

There were 52 self-reported cancer cases ascertained in the CARRS cohort. Follow-up efforts following the linkage exercise revealed that five of those were not true cancers; two participants had benign conditions based on the examination of medical records, and three participants later denied cancer diagnosis. This brought the total number of self-reported eligible cases to 47.

Of the 47 eligible cases, 29 were incident cases, 14 were prevalent cases and for the remaining 4 cases, the date of diagnosis was not known (table 1). The linkage to registry data identified 24 out of 29 self-reported incident cases (SR-RC cases) in addition to 12 NSR-RC cases, which were ascertained in registry data only. Thus, 36 incident cases were identified in the registry via linkage. The SR-RNC incident cases (n=5) were manually searched in the registry, which led to the identification of one additional case missed by the linkage, probably because of the low score. Based on these data, the self-report for incident cases compared with registry had a sensitivity of 68% (95% CI: 52% to 80%), specificity of 99.97% (95% CI: 99.91% to 99.99%), PPV of 86% (95% CI: 69% to 94%) and NPV of 99.90% (95% CI: 99.82% to 99.94%). The kappa statistic for incident cases was 0.76 (95% CI: 0.64 to 0.87).

Patients with incident cancer were older, less educated and more likely to be from lower wealth strata compared with all CARRS participants (table 2). There were no other important differences between registry-confirmed cases identified via linkage (SR-RC and NSR-RC) and the overall CARRS cohort.

The per cent agreement between CARRS and MMTR data did not differ substantially among SR-RC incident cancer cases and NSR-RC incident cases (table 3). Cases with a more recent diagnosis (2014–2015) were more likely to be ascertained via linkages in both SR-RC and NSR-RC categories. The three most common cancer sites in the SR-RC group were the gastrointestinal tract (25%), breast (21%) and head and neck (21%). The most common sites among the NSR-RC category were the gastrointestinal tract and head and neck (25% each) followed by urogenital tract and lung (17% each).

Secondary linkage for matching all (both prevalent and incident) cases for the period 1982–2015 included 11 772 individuals and 140 986 cases, respectively, in CARRS and MMTR data sets. A total of 61 cases were identified. Of those, 37 cases were SR-RC cases. The remaining 24 were

Table 1 Agreement of self-reported cancer cases in CARRS* and cancer diagnosis in MMTR†, ‡, §

Primary linkage: incident cancers from enrolment to the end of 2015¶			
	Registry data		
Self-report	**Yes**	**No**	**Total**
Yes	25**	4	29
No	12	11 726	11 738
Total	37	11 730	11 767
Sensitivity of self-report=0.68 (95% CI: 0.52 to 0.80).			
Specificity of self-report=0.9997 (95% CI: 0.9991 to 0.9999).			
PPV of self-report=0.86 (95% CI: 0.69 to 0.94).			
NPV of self-report=0.9990 (95% CI: 0.9982 to 0.9994).			
Kappa statistic=0.76 (95% CI: 0.64 to 0.87).			
Success of linkage in identifying eligible self-reported cases=82.8% (24 out of 29 self-reported eligible cases).			
Secondary linkage: all cancers (prevalent and incident) from 1982 to the end of 2015			
	Registry data		
Self-report	**Yes**	**No**	**Total**
Yes	38††	14	52
No	24	11 696	11 720
Total	62	11 710	11 772
Sensitivity of self-report=0.61 (95% CI: 0.49 to 0.72).			
Specificity of self-report=0.9988 (95% CI: 0.9980 to 0.9993).			
PPV of self-report=0.73 (95% CI: 0.60 to 0.83).			
NPV of self-report=0.9980 (95% CI: 0.9970 to 0.9986).			
Kappa statistic=0.67 (95% CI: 0.57 to 0.77).			
Success of linkage in identifying eligible self-reported cases=78.7% (37 out of 47 self-reported eligible cases).			

*Centre for Cardiometabolic Risk Reduction in South-Asia (CARRS) study.
†Madras Metropolitan Tumour Registry (MMTR).
‡A few self-reported cases with diagnosis date after 2015 treated as 'non-cases' for present analysis as MMTR data under consideration is until 2015.
§Five self-reported cases later found to be ineligible included in the self-reported numbers.
¶Excludes five self-reported cases not confirmed in registry via linkage (SR-RNC) cases (including one ineligible case) with unknown incidence/prevalence status.
**One SR-RNC manually found in registry but not via linkage.
††Includes 37 self-reported registry-confirmed (SR-RC) cases and 1 SR-RNC case manually found in registry but not via linkage.
NPV, negative predictive value; PPV, positive predictive value.

Table 2 Characteristics of overall CARRS* cohort, and CARRS participants with incident cancer confirmed in the registry via linkage during follow-up [both self-reported registry confirmed (SR-RC) and non-self-reported, registry confirmed (NSR-RC) incident cases]

	All CARRS participants (n=11 772)		CARRS patients with incident cancer (n=36)‡	
Characteristic†	**Number**	**%**	**Number**	**%**
Age (years)				
15–34	3587	30.5	2	5.56
35–54	6036	51.3	17	47.22
55+	2149	18.2	17	47.22
Gender				
Male	5436	46.2	15	41.7
Female	6336	53.8	21	58.3
Education (number of years)				
0	1068	9.1	6	16.7
1–5	2048	17.4	12	33.3
6–10	6056	51.4	15	41.7
11–23	2598	22.1	3	8.3
Occupation				
Employed	5980	50.8	15	41.6
Housewife	4745	40.3	19	52.8
Student/Retired	545	4.6	1	2.8
Unemployed	502	4.3	1	2.8
Marital status				
Single	741	6.3	0	0.0
Married	10 369	88.1	34	94.4
Widow/ widowed/ divorced	662	5.6	2	5.6
Wealth Index§ (in tertiles)				
Low	4028	34.2	18	50.0
Middle	3977	33.8	12	33.3
High	3766	32.0	6	16.7
Religion				
Hindu	9724	82.6	33	91.7
Muslim	880	7.5	1	2.7
Others	1162	9.8	2	5.6
No response/ no religion	6	0.1	0	0.0

*Centre for Cardiometabolic Risk Reduction in South-Asia (CARRS) study.
†At the time of enrolment in the study.
‡Includes all registry-confirmed cases ascertained via linkage during follow-up.
§Characterised by different household amenities (separate cooking room and toilet facilities) and assets (television, refrigerator, washing machine, microwave, mixer-grinder, mobile phone, DVD player, computer, car, motorcycle, bicycle).

NSR-RC cases; as mentioned previously, 1 SR-RNC incident case was additionally ascertained in the registry through a manual search. Of the 9 SR-RNC cases (excluding 1 SR-RNC case ascertained in the registry through manual search), 4 participants or their next-of-kin reiterated self-report of cancer diagnosis, and 5 participants were untraceable or refused to provide further information. The estimates (95% CIs) for various measures of accuracy and agreement in the analyses of all self-reported cases were 61% (49% to 72%) for sensitivity, 99.88% (99.80% to 99.93%) for specificity, 73% (60% to 83%) for PPV, 99.80% (99.70% to 99.86%) for NPV and 0.67 (0.57 to 0.77) for kappa (table 1). The corresponding estimates for self-reported prevalent cases at the time of enrolment in the CARRS study were 52% (33% to 70%) for sensitivity, 99.92% (99.85% to 99.96%) for specificity, 59% (39% to 77%) for PPV, 99.90% (99.82% to 99.94%) for NPV and 0.55 (0.38 to 0.72) for kappa (data not shown). For incident SR-RC cases, self-reported year of diagnosis was available for 9 out of 24 cases. In 6 of those 9 self-reported cases year of diagnosis was accurate. For non-incident SR-RC cases, self-reported year of diagnosis was available for 77% cases. Of those, 40% were accurate (data not shown). The registry-confirmed patients with cancer (both prevalent and incident) were more likely to be women, older, less educated and from lower wealth strata than all CARRS participants (data not shown). The agreement between CARRS and MMTR data did not differ substantially among all SR-RC and NSR-RC cases. Cancer cases diagnosed prior to 2000 were more likely to be identified exclusively via linkage (NSR-RC cases). The three most commonly self-reported cancer sites in SR-RC category were breast (27%), gastrointestinal tract (22%) and urogenital tract (19%). The sites most commonly seen among NSR-RC cases were urogenital tract (33%) and breast and head and neck (17% each) (data not shown).

DISCUSSION

While most eligible self-reported incident cases (83%; 24 out of 29) in CARRS were confirmed in the MMTR data via linkage, approximately one-third of incident cases detected by linkage in the registry would have been missed if ascertainment relied on self-report alone. It is important to acknowledge that neither self-report nor registry linkages alone can ensure complete ascertainment of all cases. Nevertheless, if the assessment relied exclusively on linkages, only 9 cases would have been missed, and even then, it cannot be stated with certainty if all of those were true cases. However, if only self-reports were relied on, 24 cases would have been missed. Especially under-reported were incident cancers of the urogenital tract and lung. Several factors may be responsible for failure to report a previous cancer diagnosis. These factors include cancer-related stigma or rapid progression of the disease, both of which may preclude accurate ascertainment during participant follow-up. It is also likely that some cases were

Table 3 Comparison of self-reported registry confirmed (SR-RC) and non-self-reported registry confirmed (NSR-RC) incident cases identified during follow-up

Variables	SR-RC cases (n=24)*	NSR-RC cases (n=12)	P value
Linkage variables (% agreement between CARRS and MMTR)			
Name	93.2	97.3	0.61
Gender	95.8	100.0	0.47
Age	93.9	97.1	0.68
Address	72.1	74.2	0.89
PIN code	98.7	100.0	0.69
Father/spouse name	67.5	72.3	0.77
Years of diagnosis (%)			
2010–2011	0.00	0.0	
2012–2013	41.7	33.3	
2014–2015	58.3	66.7	0.73
Most common primary sites (%)			
Breast	20.8	8.3	
Head and neck	20.8	25.0	
Gastrointestinal			
Tract	25.0	25.0	0.92
Urogenital tract	12.5	16.7	
Lung	8.4	16.7	
Others	12.5	8.3	

*One self-reported case manually found in registry but not via linkage excluded.
CARRS, Centre for Cardiometabolic Risk Reduction in South-Asia; MMTR, Madras Metropolitan Tumour Registry.

not captured in the CARRS cohort because they were lost to follow-up or refused to participate in the study at time points when information on "ever diagnosis" of cancer was collected. We also observed that the proportion of breast cancer cases was low in the NSR-RC group and high in the SR-RC group, indicating relatively good levels of awareness and the possible effect of screening. Further, the observation that cancer cases diagnosed before 2000 were more likely to be NSR-RC cases indicates incomplete recall. Other studies have also observed that a longer time interval between diagnosis and interview is associated with incorrect self-reporting.[2 4]

Several studies in the USA, Western Europe and parts of East Asia have linked cohort data with cancer registries, and most have used unique identifiers such as social security numbers.[4 8 23–25] Compared with our study, Inoue *et al*[24] observed substantially lower estimates for sensitivity and PPV at 53% and 60%, respectively, while assessing the validity of self-reported incident cases in Japan Public Health Center-based Prospective Study. On the other hand, Jacobs *et al*[8] successfully verified 89% of the incident cancers self-reported in the Cancer Prevention Study-3 cohort in the USA. Other studies examining the validity of self-reported history of prior cancer diagnosis have observed higher sensitivities (57.5%–87%) for prevalent cancers than observed in our study.[4 23 25]

The self-reported cancer cases not confirmed by registry linkage (ie, SR-RNC cases) in the current study warrant a closer evaluation. It is likely that some were benign or premalignant conditions such as Barrett's oesophagus or high-grade cervical dysplasia, which required diagnostic work-up and treatment, and were misinterpreted as cancer but this could not be confirmed during the follow-up efforts. At least two of the self-reported cases not found in the registry turned out to be benign conditions. Another possible explanation for failed linkage is that the residential address used at the time of hospital registration was different from the address listed in the CARRS data set. Since residence requirement in Chennai was only 1 year for CARRS participants at the time of recruitment at baseline, it is also possible that at the time of cancer diagnosis, residential addresses of participants were outside of the MMTR catchment area. Further, while MMTR maintains a comprehensive record of cancer cases, and data reliability and quality are continually monitored, it is possible that some of the eligible cases are not captured due to incomplete or erroneous records at the reporting hospitals, or because the cases were treated at hospitals outside the registry's catchment area. Finally, some of the cases may have received low scores due to discrepancies between the information in CARRS and MMTR data sets. Thus, they were not

displayed by the software, which precluded their ascertainment via linkage.

To the best of our knowledge, this is the first study in India analysing the utility of modern probabilistic linkage software for cancer research. It is especially noteworthy that the linkages in our study were carried out in the absence of unique identifiers that are readily available in HIC. For this reason, linkages had to rely on identifiers with limited discriminatory power, such as names, and addresses, which substantially increased the need for manual review. For both SR-RC and NSR-RC cases, about one-third had to be contacted via a telephone call to confirm names and addresses. However, the identification of the non-self-reported cases and the amount of information on clinical data that can potentially be captured via linkages outweigh this additional work. It is also possible that some of the true matches were missed because they had a score less than the threshold score set for manual review. While this is a potential source of linkage error, it needs to be considered in the context of feasibility, and resources available for manual review. Approximately 1700 records had to be reviewed with the current choice of threshold of ≥20. This number would have increased substantially by about 3500 records if the threshold was reduced to 15. Confirmation of unlinked self-reported cases was dependent only on the collection of additional information from the participants as lack of access rendered linkages with the vital registration system unfeasible. Another challenge encountered during the implementation of this study was the lack of consistency across data sets with respect to some of the linkage variables. Limited access to medical records restricted our ability to confirm unlinked cases. On the other hand, a high degree of data completeness in both data sets and the availability of a qualified research team to manually review the results greatly facilitated the work. When conducting manual review, the investigators took into account local naming conventions, migration patterns and residential localities.

Another important aspect that warrants consideration while conducting linkages is the ethical implications of the process and findings. Due to considerable cancer-related stigma in some population groups,[26 27] protection of privacy and autonomy of study participants are of paramount importance.

Future data linkages can benefit from lessons learnt in this study. Of the limited number of matching variables used in the present study, participant name and address (multiple sometimes) were the most important ones. Additionally, participant's age and information on additional names, that is, of father or spouse were found to be relevant. Investigators should endeavour to adopt more rigorous and standard methods of collecting information on personal identifiers and demographic variables. They should also plan on collecting information on variables with greater discriminatory power. Though subject to change, telephone numbers can provide additional identifying information. Identification numbers such as driver's licence number and voter card number can be useful. Especially promising may be the use of Aadhaar numbers, which are the 12 digit random unique identification numbers issued to the residents of India by the Government of India since 2010.[28] Basic demographic and biometric information is collected as part of the enrolment process and Aadhaar numbers are issued following deduplication and verification of the information. This unique ID system has the potential to dramatically improve both the accuracy and the efficiency of linkages and push forward population-based research in India.

Although the present pilot study included a small number of cancer cases and was restricted to one location in India, it captured patients from a wide range of sociodemographic backgrounds and yielded promising results. We conclude that probabilistic linkage methods developed and primarily applied in HIC settings can be used in LMIC such as India. The linkage efforts described in this pilot study offer cautious optimism for future research in India, especially considering data limitations, which include lack of unique numerical identifiers, inconsistency in recording addresses, and marked heterogeneity of naming conventions.

Author affiliations
[1]The Centre for Chronic Conditions and Injuries, Public Health Foundation of India, Gurugram, Haryana, India
[2]Centre for Chronic Disease Control, Dwarka, Delhi, India
[3]Cancer Institute-WIA, Chennai, Tamil Nadu, India
[4]Genentech Inc, South San Francisco, California, USA
[5]Madras Diabetes Research Foundation (ICMR Center for Advanced Research on Diabetes), Chennai, Tamil Nadu, India
[6]Ramaiah International Centre for Public Health Innovations, Bengaluru, Karnataka, India
[7]Centre for Health, Innovation and Policy, Noida, Uttar Pradesh, India
[8]Department of Epidemiology, Rollins School of Public Health, Atlanta, Georgia, USA
[9]North American Association of Central Cancer Registries, Springfield, Illinois, USA
[10]Dr. Mohan's Diabetes Specialities Centre (IDF Centre of Excellence in Diabetes Care), Gopalapuram, Chennai, Tamil Nadu, India
[11]Department of Hematology & Medical Oncology, Emory University School of Medicine, Atlanta, Georgia, USA
[12]Emory University Winship Cancer Institute, Atlanta, Georgia, USA
[13]Department of Population Science, American Cancer Society, Atlanta, Georgia, USA
[14]Centre for Cancer Statistics, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA

Acknowledgements We gratefully acknowledge Centre for Cardiometabolic Risk Reduction in South-Asia (CARRS) study participants for giving their consent for linking their data with disease registries. This work was made possible by the extensive support of the field team of CARRS study, in particular Kumar Munusamy, Vel Murugan and Shobana. We are also thankful for the support received from the Madras Metropolitan Tumour Registry team. AA was supported by DBT/Wellcome Trust India Alliance Fellowship (Grant number: IA/CPH /17/1/503340) at the time of data analysis and drafting of the manuscript.

Contributors Specifically, the authors made the following contributions: AA: Substantial contributions to the design of the work; acquisition, analysis and interpretation of data for the work, drafting the manuscript. AA is also the guarantor for the overall content of this article. RR: Substantial contributions to the design of the work; acquisition, and interpretation of data for the work. PKD: Substantial contributions to the conceptualisation and design of the work; acquisition, analysis and interpretation of data for the work. MD: Substantial contributions to the design of the work; acquisition, analysis and interpretation of data for the work. DK: Substantial contributions to the acquisition, analysis and interpretation of

data for the work. NK: Substantial contributions to the acquisition, analysis and interpretation of data for the work. DB: Substantial contributions to the design of the work; acquisition of data for the work. RM: Substantial contributions to the conceptualisation of the work. BK: Substantial contributions to the conceptualisation and design of the work. VM: Substantial contributions to the design of the work; acquisition and interpretation of data for the work. TWG: Substantial contributions to the conceptualisation and design of the work. AP: Substantial contributions to the conceptualisation and design of the work, analysis and interpretation of the data. SR: Substantial contributions to the conceptualisation and design of the work; acquisition and interpretation of data for the work. DP: Substantial contributions to the conceptualisation and design of the work; acquisition and interpretation of data for the work. KW: Substantial contributions to the conceptualisation and design of the work; analysis and interpretation of data for the work, drafting the manuscript. MG: Substantial contributions to the conceptualisation and design of the work; analysis and interpretation of data for the work, drafting the manuscript, AND all authors met the following criteria: (1) Revising the work critically for important intellectual content; AND (2) Final approval of the version to be published; AND (3) Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding We acknowledge the funding support provided by the National Cancer Institute (NCI), National Institute of Health, USA, (Grant Number: P20CA210298) to carry out this work.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This work was approved by local Institutional Ethics Committees of Centre for Chronic Disease Control (CCDC), Delhi, India (CCDC-IEC_08_2017) and Madras Diabetes Research Foundation (MDRF), Chennai, India. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The data underlying this article (personal identifiers of the CARRS study participants and individuals in MMTR database) cannot be shared publicly because of privacy issues. Data will be shared on reasonable request to the principal investigators of the two data sources—DP at dprabhakaran@phfi.org for CARRS study and RS at r.swaminathan@cancerinstitutewia.org for MMTR.

ORCID iDs

Aastha Aggarwal http://orcid.org/0000-0002-8180-8743
Theresa W Gillespie http://orcid.org/0000-0001-8734-716X
Michael Goodman http://orcid.org/0000-0001-6956-6879

REFERENCES

1 Bergmann MM, Calle EE, Mervis CA, *et al*. Validity of self-reported cancers in a prospective cohort study in comparison with data from state cancer registries. *Am J Epidemiol* 1998;147:556–62.
2 Desai MM, Bruce ML, Desai RA, *et al*. Validity of self-reported cancer history: a comparison of health interview data and cancer registry records. *Am J Epidemiol* 2001;153:299–306.
3 Koller KR, Wilson AS, Asay ED, *et al*. Agreement between self-report and medical record prevalence of 16 chronic conditions in the Alaska earth study. *J Prim Care Community Health* 2014;5:160–5.
4 Navarro C, Chirlaque MD, Tormo MJ, *et al*. Validity of self reported diagnoses of cancer in a major Spanish prospective cohort study. *J Epidemiol Community Health* 2006;60:593–9.
5 Kool M, Bastiaannet E, Van de Velde CJH, *et al*. Reliability of self-reported treatment data by patients with breast cancer compared with medical record data. *Clin Breast Cancer* 2018;18:234–8.
6 Bernstein L, Allen M, Anton-Culver H, *et al*. High breast cancer incidence rates among California teachers: results from the California teachers study (United States). *Cancer Causes Control* 2002;13:625–35.
7 Bisgard KM, Folsom AR, Hong CP, *et al*. Mortality and cancer rates in nonrespondents to a prospective study of older women: 5-year follow-up. *Am J Epidemiol* 1994;139:990–1000.
8 Jacobs EJ, Briggs PJ, Deka A, *et al*. Follow-Up of a large prospective cohort in the United States using linkage with multiple state cancer registries. *Am J Epidemiol* 2017;186:876–84.
9 Bergmann MM, Noethlings U, Eisinger B, *et al*. The importance of the common cancer Registry for the identification of cancer cases in the EPIC potsdam-study -- results of the first record linkage. *Gesundheitswesen* 2004;66:475–81.
10 Obi N, Waldmann A, Babaev V, *et al*. Record linkage of a large clinical practice patient cohort with the cancer registry Schleswig-Holstein. *Gesundheitswesen* 2011;73:452–8.
11 Sengayi M, Spoerri A, Egger M, *et al*. Record linkage to correct under-ascertainment of cancers in HIV cohorts: the sinikithemba HIV clinic linkage project. *Int J Cancer* 2016;139:1209–16.
12 India State-Level Disease Burden Initiative Cancer Collaborators. The burden of cancers and their variations across the states of india: the global burden of disease study 1990-2016. *Lancet Oncol* 2018;19:1289–306.
13 NCDIR-NCRP. Report of national cancer registry programme (2012-2016. 2021. Available: https://ncdirindia.org/All_Reports/Report_2020/resources/NCRP_2020_2012_16.pdf
14 India State-Level Disease Burden Initiative Collaborators. Nations within a nation: variations in epidemiological transition across the states of india, 1990-2016 in the global burden of disease study. *Lancet* 2017;390:2437–60.
15 Nair M, Ali MK, Ajay VS, *et al*. CARRS surveillance study: design and methods to assess burdens from multiple perspectives. *BMC Public Health* 2012;12:701.
16 NCDIR-NCRP. Population based cancer registry, chennai. cancer institute (WIA), adyar, chennai. 2021. Available: https://www.ncdirindia.org/All_Reports/PBCR_REPORT_2012_2014/ALL_CONTENT/PDF_Printed_Version/Chennai_Printed.pdf
17 CensusInfo india 2011. 2021. Available: http://www.dataforall.org/dashboard/censusinfoindia_pca/
18 District census handbook, chennai. 2021. Available: http://censusindia.gov.in/2011census/dchb/DCHB_A/33/3302_PART_A_DCHB_CHENNAI.pdf
19 Indian names. 2021. Available: https://www.behindthename.com/glossary/view/indian_names
20 Indian name. 2021. Available: https://en.wikipedia.org/wiki/Indian_name
21 Match*Pro software. 2021. Available: http://surveillance.cancer.gov/matchpro/download/75756-CcntAe1io2
22 Ali MK, Bhaskarapillai B, Shivashankar R, *et al*. Socioeconomic status and cardiovascular risk in urban South Asia: the CARRS study. *Eur J Prev Cardiol* 2016;23:408–19.
23 Cho S, Shin A, Song D, *et al*. Validity of self-reported cancer history in the health examinees (HEXA) study: a comparison of self-report and cancer registry records. *Cancer Epidemiol* 2017;50:16–21.
24 Inoue M, Sawada N, Shimazu T, *et al*. Validity of self-reported cancer among a Japanese population: recent results from a population-based prospective study in Japan (JPHC study). *Cancer Epidemiol* 2011;35:250–3.
25 Nash SH, Day G, Hiratsuka VY, *et al*. Agreement between self-reported and central cancer registry-recorded prevalence of cancer in the alaska earth study. *Int J Circumpolar Health* 2019;78:1571383.
26 Nyblade L, Stockton M, Travasso S, *et al*. A qualitative exploration of cervical and breast cancer stigma in karnataka, india. *BMC Womens Health* 2017;17:58.
27 Gupta A, Dhillon PK, Govil J, *et al*. Multiple stakeholder perspectives on cancer stigma in North India. *Asian Pac J Cancer Prev* 2015;16:6141–7.
28 India UIAo. AADHAAR: unique identification authority of india. 2021. Available: https://uidai.gov.in/my-aadhaar/about-your-aadhaar.html