SIA | SOCIAL INVESTMENT AGENCY ORANGA TANGATA

Investing in what works for better lives

Representative timelines – modelling people's life experiences

November 2019

Analytic methodology

SIA-1911.DP.MET.WEL

New Zealand Government

Creative Commons Licence

Integrated Data Infrastructure disclaimer

The results in this paper are not official statistics, they have been created for research purposes from the Integrated Data Infrastructure managed by Statistics New Zealand. The opinions, findings, recommendations and conclusions expressed in this paper are those of the author(s) not Statistics NZ, or other government departments.

Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation and the results in this paper have been supressed to protect these groups from identification.

Careful consideration has been given to the privacy, security and confidentiality issues associated with using administrative and survey data in the Integrated Data Infrastructure. Further detail can be found in the Privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form, or provided to Inland Revenue for administrative or regulatory purposes.

Any person who has had access to the unit record data has certified that they have been shown, have read, and have understood section 81 of the Tax Administration Act 1994, which relates to secrecy. Any discussion of data limitations or weaknesses is in the context of using the Integrated Data Infrastructure for statistical purposes and is not related to the data's ability to support Inland Revenue's core operational requirements.

Liability

While all care and diligence has been used in processing, analysing and extracting data and information in this publication, the Social Investment Agency gives no warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information in this publication.

Attribution

Author: Simon Anastasiadis

Project team: Simon Anastasiadis†, Belinda Gorman‡, Thalia Wright‡, Karen Clifford‡, Akilesh Chokkanathapuram†, Vicki Evans†, Michael Hackney†, Peter Holmes†, Athira Nair†, Justine Stephen†, Fiona Thomson†, Kate Kolich†

Consulted: Gael Surgenor‡, Luella Linaker‡, Matthew Spencer, Marianna Pekar†, Atawhai Tibble†, Lee O'Connor†, Danny Mollan†

† = Social Investment Agency. ‡ = The Southern Initiative.

Special thanks: The Southern Initiative, Stats NZ Integrated Data Team

Citation

Social Investment Agency 2019. Representative timelines – modelling people's life experiences. Wellington, New Zealand.

ISBN 978-0-473-50410-6 (online)

Published in November 2019 by
Social Investment Agency
Wellington, New Zealand

Contents

Foreword from The Southern Initiative

Collaboration, innovation, and learning more about what really works to help whānau and communities thrive are at the heart of both The Southern Initiative and The Social Investment Agency.

Given this strong commonality, TSI was keen to partner with SIA to see if there were ways in which we could combine the data access and expertise that the Social investment Agency has with our whānau-centric, place-based kaupapa to strengthen the effectiveness and impact of both organisations' work.

From the beginning we were asking two questions. The first was a process question: Can we find ways of making big data accessible and useful to whānau – and just as importantly, make whānau experience accessible and useful in the design and meaning making of big data projects? The second was an outcomes question: Can the combination of our different skill sets help to unlock new insights and find new areas for action that could improve outcomes for whānau?

For TSI the answer to both questions was 'yes'. The value of the partnership work with SIA has included:

- further building our evidence base of the weight and specific shape of toxic stress which whānau in South Auckland are experiencing
- demonstrating – by using existing data – the complexity of people's lives to build empathy in new ways
- highlighting new areas for action and investigation – especially around the work patterns of fathers and the value of informal post-natal support
- recognising how much whānau valued engaging with data and the positive benefits of democratising data rich conversations.

– Gael Surgenor, Director - Community and Social Innovation

Representative timelines provide a starting point for understanding

Understanding the life experiences of people and families is the focus of research throughout the social sector. With the increasing availability and granularity of data — especially administrative information — more detailed and nuanced analytics are becoming possible. Representative timelines are a tool to create insight from people's interactions with social services. They can provide a starting point for determining what questions might help us gain better understanding of people's experiences.

The Social Investment Agency (SIA) has completed an analysis using cross-sector representative timelines of families' experiences. This work has shown representative timelines to be effective at drawing information together to generate new questions and insights.

A timeline is a sequence of potentially overlapping events and periods which reflect significant parts of people's life experiences. A representative timeline seeks to capture the experiences that are common across a group of people — similar to a group average.

Representative timelines can generate new insight by bringing together different aspects of people's experiences into one consistent view — allowing researchers to see the breadth of events and the interactions between those events, over time. The 'representative' aspect of this approach also provides confidentiality, protecting the identity of the people whose data is used. The 'timeline' aspect of this approach combines well with lived experience feedback as people's stories add context and meaning to the timeline.

This document provides guidance to support others in the construction of representative timelines. Examples drawn from the first application of this work are used throughout this document to demonstrate the approach.

This technique was developed in a study of the journey of families having a baby

This technique was developed, and first applied, for a study of South Auckland families' experiences around the birth of a child. This work was completed in partnership between SIA and The Southern Initiative (TSI).

The Southern Initiative is a place-based initiative that stimulates, enables and champions social and community innovation across the Local Board areas of Papakura, Manurewa, Ōtara-Papatoetoe, and Māngere-Ōtāhuhu. As a part of Auckland Council, TSI works to find radical solutions to some of South Auckland's most pressing social and economic challenges.

Partnering with TSI helped to ensure that the information used to create insights was relevant and described real experiences. TSI's rich understanding of the South Auckland context guided the technical research and facilitated the inclusion of people's stories and lived experiences alongside the analytics.

The experience of families in South Auckland about the birth of a baby, and its impact on the wellbeing of the baby, was of significant interest to TSI. Prior to beginning our partnership, they had reviewed the science of babies' brain development. Their review highlighted that parents under high burdens of stress struggled to have the mental resources needed to parent well, with detrimental impacts on the developing child (The Southern Initiative & The Auckland Co-Design Lab, 2017).

Informed by some initial conversations with parents in South Auckland, TSI were keen to see whether sources of stress could be observed in the data. They also hoped to discover possible points for intervention, and to identify questions to investigate further. Representative timelines were agreed as an approach to connect the science, the data, and the lived experience.

SIA constructed representative timelines of families' journeys from nine months before the birth until six months after between 2005 and 2017. Representative timelines were constructed for the baby, both parents, and any siblings, where events could be inferred from data in the Integrated Data Infrastructure (IDI). Similar timelines and experiences were grouped together, and timelines of representative experiences were used to generate insights.

The intention of using representative timelines was not just to provide an analytic study but also to make the outputs relatable, so that the people who are represented in the timelines might be able to see themselves reflected in the data. There was strong agreement in the partnership that we wanted to understand the people, not just the data about them. The representative timelines provided a starting point from which TSI went on to engage further with whānau, enabling create a deeper understanding of their lived experience.

The application of this technique aligned with the Data Protection and Use Principles[1]

SIA is committed to the safe, ethical and transparent use of social sector data. We have worked with the social sector to collectively develop a policy for anyone working with personal data and information (including information that has been de-identified or anonymised) [2].

The Data Protection and Use Policy (DPUP) is a collection of Principles and guidance that will be supported by tools to enable everyone to easily understand what is appropriate, what is not, and how to work respectfully with people's personal information. DPUP will help to build trust and confidence in the social sector's use of people's data and information, and the quality of that data so that better outcomes can be achieved from it. It is important to know that this policy promotes best practice but will not change legislation.

For our initial application of the representative timeline approach, our partnership with TSI was important in how we aligned with the Principles of DPUP:

- TSI had engaged with families and service providers in South Auckland before any data preparation for representative timelines began. The things they learned from these

[1] Given the timing of this work, we used the draft Policy Principles.

[2] https://www.sia.govt.nz/investing-for-social-wellbeing/data-protection-and-use/

conversations guided SIA staff as they developed and applied the representative timeline technique.

- Both SIA and TSI were committed from the beginning to sharing our findings with whānau in South Auckland. Seven early insights from the timelines were used as a starting point for conversations with whānau following the completion of the analytics. This let them see how their data was being used and created an opportunity for us to hear their experiences. Each of the early insights was reconsidered and reinterpreted in the context of what whānau told us.
- Following the conclusion of the research on South Auckland families having a baby, representative timelines and feedback from the whānau engagement were consistently presented together.

In these ways, SIA is confident that the transparency and inclusivity with which we first conducted representative timeline research upholds the Principles set out by DPUP. Further examples arise in the following sections on data environment and timeline interpretation.

The technique has analytic strengths and weaknesses

From our first application, we have observed representative timelines to be an effective analytic technique for three main reasons: They combine data from a wide range of sources, they encourage empathy, and they can be extended with additional qualitative or quantitative information.

Almost all data relating to people has a time component. As timelines combine information based on its timing, even very different data collections can be included on a common timeline. This enables timelines to provide a picture that crosses organisational or departmental boundaries, encouraging decision makers to consider the impact of organisations other than their own, broadening the range of available options, and providing a point around which to collaborate.

The objectiveness and cerebral nature of traditional statistics makes them well suited to providing their readers with intellectual understanding. However, these same features can also dehumanise the people they describe. Because human experience is sequential through time, and timelines convey these sequences, a timeline can help the viewer appreciate what it would be like to live the journey. This helps the reader engage with their empathy, not just their intellect. These different ways of engaging with timelines can lead to deeper understanding and help make them accessible to readers for whom traditional statistics are problematic.

Not all information, or understanding, can be well represented on a timeline. However, an analytic timeline provides a strong starting point from which further information can be gathered and added. Where areas needing further investigation are identified, qualitative information, such as case studies or ethnographic research, can be gathered and specific points or windows of time related to a life event can be chosen for more detailed quantitative study. These can then be connected back to the timeline, without which it would be difficult to combine such information.

In addition to the above strengths, a further advantage of representative timelines is that they provide a way of protecting the confidentiality of the people whose data is used. By combining individuals to create a synthetic representative timeline, analysis can ensure that an average experience is available for examination, without revealing any single individual's experience.

Timelines are not without their share of limitations. Just as established statistical techniques are fit for some purposes and not for others, so it is with timelines. The main limitations we observed were restrictions around what can be shown, and imprecision in how the representative aspect of a timeline is understood.

Timelines are well suited to show the occurrence, and non-occurrence, of events and periods. However, they are not well suited to show more continuous measures across time. This means that the level or intensity of an event can be difficult to display. We encountered this in our application to South Auckland when considering income from employment: We could show occurrence (or not) of income over a certain threshold, but income as a continuous measure could not be directly included on the timelines. It was instead necessary to use additional plots to provide this information.

Representative timelines are often more difficult to interpret than individual timelines. This is in part because what makes a timeline representative is not as well defined or intuitive as other, more traditional, summary measures such as averages. The consequence of this is that additional effort is required to understand the reality that is communicated by a representative timeline. This will become less of a concern as these techniques see more use and refinement.

Our tools are available to support constructing other timelines

The techniques to construct timelines can be applied to a range of other situations. SIA have developed a collection of resources that can be used for investigating other timelines. In addition to this guidance document there is data preparation code, a visualisation tool, and the documentation for both. We have sought to keep these resources general, so they are applicable to both other types of journeys and other data sources.

The code for both the data preparation and the visualisation tool is available on SIA's GitHub page[3]. This consists of:

- The repository for the data preparation code[4] covers the organisation of the raw data into individual level timelines before combining it into representative timelines.
- The repository for the visualisation tool[5] covers the creation of an interactive timeline visualisation app to support the exploration of the data.

SIA makes a practice of releasing code associated with our analytical projects once the project is complete. Please contact us by email (info@sia.govt.nz) if you want to know more.

[3] https://github.com/nz-social-investment-agency

[4] https://github.com/nz-social-investment-agency/representative_timelines

[5] https://github.com/nz-social-investment-agency/timeline_visualisation

Timelines could be extended in a variety of ways

The focus of our initial project was to demonstrate the validity and usefulness of representative timelines as an analytical tool. As a result, there are a range of refinements and extensions that we are yet to explore. The following list covers several such ideas we believe merit further consideration:

- The current method for producing representative timelines seeks to preserve timing relative to the reference date. Techniques that preserve other details, such as sequences of events, duration of events, or gaps between events need to be considered to meet other research needs.
- Timelines can imply causal association, but without accompanying analytics the strength of such implications can not be determined. The addition of techniques that identify interactions between pairs of events enable analysis to make statements of the form: Of people who have experience X, Q% of them had experience Y beforehand, and went on to have experience Z after.
- At present, our method for constructing representative timelines requires access to the underlying individual timelines. There is a need for techniques that let us combine representative timelines directly and produce the same global representative timeline as using the pooled individual timelines. This would enable more dynamic analyses and motivate the development of more robust mathematics for constructing representative timelines.
- Our present techniques for identifying groups of similar timelines do not extend easily to include non-timeline information (for example, clustering by timeline and total income). Generalising these techniques would allow for different types of clustering, identifying different kinds of patterns in the data.

Timelines are not the best solution for every application

The rest of this paper provides guidance for others who are seeking to create their own representative timelines. However, it is important to first consider whether representative timelines are the best solution as other techniques exist for understand sequences of events through time.

Several distinctions merit consideration:

1. Whether multiple events can occur simultaneously. Modelling an experience where events occur sequentially and cannot overlap (for example: no benefit, applies for benefit, receives benefit) is different from modelling an experience where events can occur concurrently (for example: being employed, enrolled in study, married).
2. Whether the repetition of events is important. Modelling an experience that iterates through the same set of events (for example: healthy, sick, recovering, healthy, sick) is different from modelling an experience where events follow one-another linearly (for example: appointment scheduled, checking in, waiting, having appointment, checking out).

3. Whether events can be of different durations. Modelling an experience where every event has the same duration, or duration is unimportant is different from modelling an experience where there are significant differences in the duration of events.

Given these distinctions, several other techniques deserve mention:

- Sankey or alluvial diagrams are well suited to examining situations where events do not occur concurrently. While they can be adapted to consider the repetition of events, or different event durations, this is not their focus.
- Network analysis including Markov chains and transition matrices, are well suited to examining situations where the repetition of the same set of events is of interest. While they can be adapted to consider multiple concurrent events, or different event durations, this is not their focus.
- Relationship timelines are effective for analysing time-varying network data. This enables them to consider multiple concurrent events and different durations. While they can be adapted to consider the repetition of events or non-network data, this is not their focus.

In contrast to these, representative timeline modelling works well when multiple events can occur concurrently, and the duration of events may differ. It focuses on events over time and hence is not well suited for cyclic analyses where the focus is on a repeated sequence of events.

Clarity of scope is required to define the timelines

The scope of timelines determines their contents and how well they fulfil their purpose. While some exploration and iteration are common in any analytics undertaking, establishing the scope of what the timelines are intended to cover focuses the analysis and enables it to proceed with confidence.

Timelines can be defined by four sets of decisions: who, when, what, and where. These decisions should be preceded and informed by "why": the motivation for constructing timelines. Once these decisions are made, focus turns to "how": the methodology for construction and visualisation.

Who – the people that will be included

The people who will be included in the analysis need to be determined. Often, the inclusion of an individual depends on some mixture of their attributes and their experiences. For example: in a study of school leavers, a researcher could restrict their study population to domestic students (an attribute) who leave with a high school qualification (an experience).

As part of specifying who will be included, some reference time must also be provided. This time is used to synchronise across different people and is often the date of the experience all the people have in common. It follows that the same person, or people, may be included multiple times in the study if they have an experience more than once. For example: In a study of patient experiences, a person could be included once for each time they are a patient at a hospital, with the date of

admittance as the reference time. However, while the individual may be the same, the reference time will be distinct.

There is no requirement that every person appears on a distinct timeline. Groups of people with a common experience can share the same timeline. For example: a timeline could be constructed for an entire whānau by connecting the timelines of individual members.

Combining these, the 'who' component of a single timeline is defined by three pieces of information: The identity of the individual, the reference time and a label for the group that the individual shares their timeline with.

For our application to South Auckland, we began with babies born to mothers who's address at the time of birth was within one of the four local board areas that make up South Auckland. Using birth records, we connected the new-born to the rest of their family: both parents (where recorded), and the children of either parent. These are the identities for which timelines were built. Adoptions and surrogate parents are sufficiently rare that they could be included without special treatment for the analysis, so were deemed out of scope.

The role of each identity was described relative to the baby (mother, father, baby, full and half sibling) and a common label was given to all members of a family to ensure they would share a timeline. The birth date of the baby was used as the reference time.

South Auckland is known to contain a significant Maori and Pacific population for whom whānau, or extended family, is important. One of the limitations of the initial study was that identifying extended family and social networks, such as whānau, was not possible with the available data.

When – the time periods for analysis

While the reference date provides an initial point from which to specify the timeline, further decisions are needed. First among these decisions is the length and resolution of the timeline.

The length of the timeline, and its position relative to the reference date, needs to be decided. For some applications a timeline of variable length may appear to better fit the research question. However, this introduces significant complexities when combining individual timelines to form representative timelines. Therefore, a timeline of fixed length is recommended for this application. For example: a study of tertiary students may wish to define a timeline from first enrolment until final graduation. But some students will graduate in three years, while others will take five or longer. Considering all students for five years, starting from the date of their enrolment, removes the complexity of different lengths of timelines.

The resolution of the timeline is the accuracy with which it measures time. In general, longer timelines are created at coarser resolutions. For example: if the length of the timeline is a few hours then measuring the timing of events to the nearest minute or ten minutes would be effective, but if the length of the timeline was a year then a weekly resolution would be more suitable.

One consequence of deciding the resolution of the timeline is that periods can be defined, and events can be recorded, by the periods of the timeline they occur in. While it is common for the source data to record events by their start and end times, converting this to periods at the

timeline resolution is recommended for summarising and confidentialising the timeline. For example: in a timeline at a weekly resolution it is preferable to know an individual was employed in weeks two, three and four instead of the exact dates of their employment.

For our application to South Auckland, we consider a timeline 15 months long: starting nine months prior to the date of birth and ending six months following the birth. Fortnightly resolution was used, dividing the timeline into 33 fortnights, labelled from -20 to 13, with the birth at time zero.

What – the measures of interest

The content that will fill the timeline: the measures of interest need to be specified. While these are often straightforward to describe, there are several additional details to consider.

First, not all measures will reflect the same type of thing. A distinction may be necessary between a point in time event (for example, a change in address), and a spell or period (for example, time spent at an address). For some measures, what form they take will depend on the resolution of the timeline (for example, a doctor's visit is a spell when measured in minutes, but a point in time when measured in days). Depending on the application different measures may need to be treated differently.

Second, whether and how source records will be combined to define measures needs to be determined. Some measures may require records from multiple sources. For example: patient attendance from referrals may require combining records from attendance and referral sources.

Another common situation for this to arise is where multiple records appear in the data that represent a single experience for a person. Whether these records are combined or not depends on who is the subject of the timeline. For example: lab test data may capture each test that was conducted following a single interaction with a patient. Hence for a patient-centric timeline combining these records to a single event for the occurrence of any lab tests would be recommended.

For our application to South Auckland, we drew upon a variety of events from IDI records. As we were interested in person-centric timelines, all measures were expressed as indicators that an individual had a specific experience at a given time. Defining measures in this way did discard some information about the intensity or magnitude of the event. For example: ACC data records both the occurrence of an accident and provides indications of its seriousness. However, this simplification was consistent with the purpose of the study.

For simplicity, measures for both point in time events and spells were prepared in the same way. A list of measures used in our initial study can be found in the appendix.

Where – the choice of data environment

The choice of data environment, where the data will be drawn from, must be made in tandem with the decisions of who, when and what. For many applications this choice will be made on the practical basis of what sources of data does the researcher have access to? However, access is not sufficient by itself to justify the use of a data source. Consideration must also be given to:

- the suitability of the contents of the environment – is it up to date, sufficiently detailed, reliable, and does it capture the measures that are wanted in the timelines? (Social Investment Agency, 2018a).
- the purpose for which the data was gathered, the permissions attached to it, and how the analysis respects the people whose data are used (Social Investment Agency, 2018b).

For our application to South Auckland, we made use of the Integrated Data Infrastructure (IDI). The IDI is a large research database assembled and maintained by Statistics New Zealand. It contains deidentified microdata from a range of government organisations. Consideration was given to the suitability and appropriateness of using the data for this analysis:

- Measures were drawn from a range of datasets where SIA staff already had significant experience to be confident of the data quality. Where staff were less familiar with the datasets required, documentation was consulted and checks for consistency were carried out before new measures were constructed.
- Following the production of the initial output, SIA staff met with representatives from the organisations who collected the data to ensure our use of it was consistent with their expertise.
- As some IDI datasets are updated less frequently than others, the analysis was focused on the period when all the required data sources were available and considered reliable.
- As part of respecting the people whose data was used, we were diligent in ensuring processes around privacy and confidentiality of the IDI were followed (Stats NZ 2016 & 2017). This included access to the IDI only under the 'Five Safes' framework, and discussions with Stats NZ prior to beginning work, to ensure the safety of our intended approach.
- Much of the data in the IDI is administrative – produced by business activity rather than for research purposes. An awareness of the limitations of depending on administrative data informed the ways we drew conclusions. That the project team included both staff from TSI who knew whānau in South Auckland and experienced IDI users from SIA, ensured that all insights were informed by both relational and technical considerations.

Given these considerations, and the lack of an alternative cross-sector data environment, we are confident that the IDI was a suitable choice of data environment for the analysis.

We have developed a repeatable process for creating representative timelines

SIA has developed a methodology for creating representative timelines. This methodology is flexible to a range of decisions about the scope of the timelines described above. Although it has been developed in the context of the IDI, the approach is applicable in general.

This section focuses on the data manipulation and processing for creating timelines. The process was developed in two distinct parts:

1. preparation of a dataset at an individual resolution, and

2. condensing of individual timelines to a smaller number of representative timelines that preserve the privacy and confidentiality of individuals.

This division provides an intermediate point for investigation and quality assurance. Due to the exploratory nature of our first application of this process, the halfway point also enabled us to confirm that meaningful timelines could be built. Researchers wanting to produce only individual level timelines will want to focus on just the first half of the process. Flow diagrams for both parts can be found in the appendix.

The techniques described in this section have well established foundations in sequence analysis. Developed in bioinformatics for analysing genetic sequences and subsequently adopted into the social sciences (Abbott, 1995), sequence analysis is a collection of techniques for working with ordered series of data. For an introduction to the theory of sequence analysis we recommend Giegerich et al. (2012); for its application to the social sciences we recommend Gabadinho et al. (2010) and the references contained within.

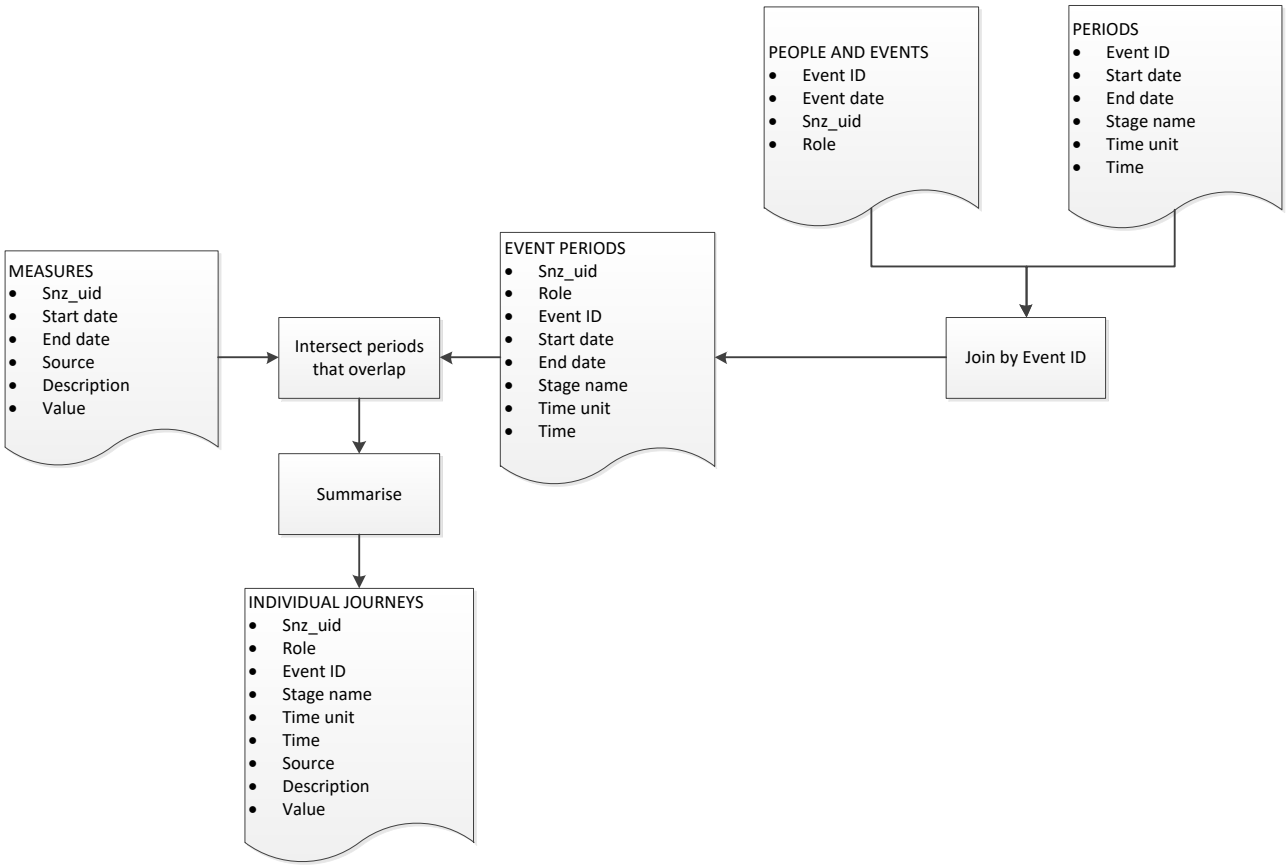
For our analysis a timeline may be considered a special type of sequence. Though the clustering of timelines and the construction of representative timelines may seem unusual, these techniques already exist for sequences (see for example Mika & Rost 2003, Gabadinho et al. 2010, Fu et al. 2012). We have applied variations that are applicable to our case.

Input measures are mapped onto time periods to produce individual timelines

The core of the analysis required for preparing individual timelines is to convert the data from its raw format into sequences describing the occurrence of events. This is done by defining the measures and the timeline periods for each individual, intersecting the two, and summarising the overlap.

When timelines are concerned only with the occurrence or absence of events, the desired summary is often an indicator for whether any records of the specified type occurred in each period. This gives a single value for each individual, for each measure, for each period. The following figure provides an overview of this process. A more extensive flow diagram recording our methodology is given in the appendix.

Figure 1: Interaction of measure and time periods to produce individual timelines



For our application to South Auckland, every measure was summarised to produce an indicator for whether any records of the specified type intersected with each fortnight period. This means that multiple records of the same type were combined to a single measure where they shared the same fortnight. It also means that an event bridging two adjacent periods can be reported as occurring in both periods. This is easiest to observe in our study where mothers who gave birth in hospital commonly have a hospitalisation event for both the period immediately before, and the period immediately after the birth.

In the IDI, data is stored in a variety of forms, most often depending on the process by which it was collected. As part of our best practice, we implemented a layer of abstraction between the raw database and our analysis to ensure consistent data preparation and formatting. This intermediate layer improved the speed of our analysis, especially when updates to the inputs or methodology were required.

Another part of our best practice was in the use of control files to simplify specifying the parameters of the data preparation. Determining which measures should be constructed for which roles (e.g. only mothers should have maternity measures) and how these measures should be prepared was specified in one control file, while the definition of each period relative to the date of birth was specified in another. A further advantage of these control files is that together they document much of the data input.

Groups of similar timelines are identified

While some applications will make use of individual resolution timelines, most research will have more individual timelines than it is practical to review. Grouping and combining timelines provides an effective way for researchers to draw conclusions from a collection of timelines. In addition, by grouping individual timelines together the confidentiality of the people whose data is used can be preserved.

When constructing groups of timelines, it is important that the members of the groups are similar so that conclusions drawn from the group are representative of all its members. There are two broad approaches to constructing these groups: manual, according to characteristics specified by researchers (for example, grouping together people based on ethnicity); and algorithmic, using computational techniques to identify clusters of similar journeys. The user defined approach enables researchers to consider journeys for groups that are already known. The computer defined approach enables researchers to discover the common kinds of journeys.

Algorithmic clustering depends on computation methods for determining the similarity or difference between a pair of sequences. Sequence analysis provides a range of suitable methods. Gabadinho et al. (2010) recommend optimal matching, introduced to social sciences by Abbott & Forrest (1986). Applying this method to each measure in the timelines enables the calculation of the similarity between all pairs of timelines. With the calculation of similarity complete, a researcher may then apply their preferred clustering algorithm.

For our application to South Auckland, both user and computer defined groups were used. User defined groups were constructed where there were known characteristics of interest (such as mothers' age or babies' birth weight). Computer defined groups were constructed by a clustering algorithm that sought to gather together individuals with similar journeys (for example it identified a group where the mothers are working and then take paid parental leave).

For the algorithmic clustering the optimal matching algorithm was chosen for the similarity measure as it is well established and recommended. Clustering of timelines was done using K-medoid clustering (also known as: Partitioning Around Medoids). A medoid is a central example in a cluster, in a similar way to how a median is a central point in a distribution.

The groups produced using the algorithmic clustering are mutually exclusive: each individual timeline was included in exactly one cluster. The manual groups are not mutually exclusive: the same individual timeline can be included in multiple groups.

For our application, memory limits in the IDI R programming environment prevented us from clustering all the timelines at once.[6] In response to this we used batch- and meta-clustering:

1. The collection of timelines was randomly divided into a series of batches, and each batch was clustered separately. If each random batch is representative of the entire collection, then each batch should produce similar types of clusters.

[6] The R programming environment is a shared resource with at most 150 GB of memory available, shared between all users. A full naïve clustering would have required more than 2000 GB of memory, much more than was available. Using the meta-clustering, we kept our instantaneous memory use below 10 GB, respecting others' use of the shared environment.

2. The medoids from each cluster were extracted from each batch. Each medoid can be considered as a representative of its cluster.
3. Clustering was run on the collection of medoids (a meta-clustering), and the results propagated back to the individual batches. The final cluster for a timeline is the cluster its medoid was assigned in the meta-clustering.

Similar timelines are condensed

Given groups of similar timelines, a representative timeline can be produced for each group. There are many ways this can be accomplished; the most appropriate approach for a specific application will be broadly determined by the following three questions:

1. Do the representative timelines need to protect the confidentiality of the data?
2. To what extent do interactions between different measures need to be preserved?
3. What features of the timeline are most important (for example: duration of events, timing relative to reference date, timing relative to other events, clustering of similar events)?

Where confidentiality is a concern a model or averaging process will need to be considered. This will require additional analytic effort. However, if confidentiality is not a concern, then a single individual resolution timeline could be chosen as the representative of its group.

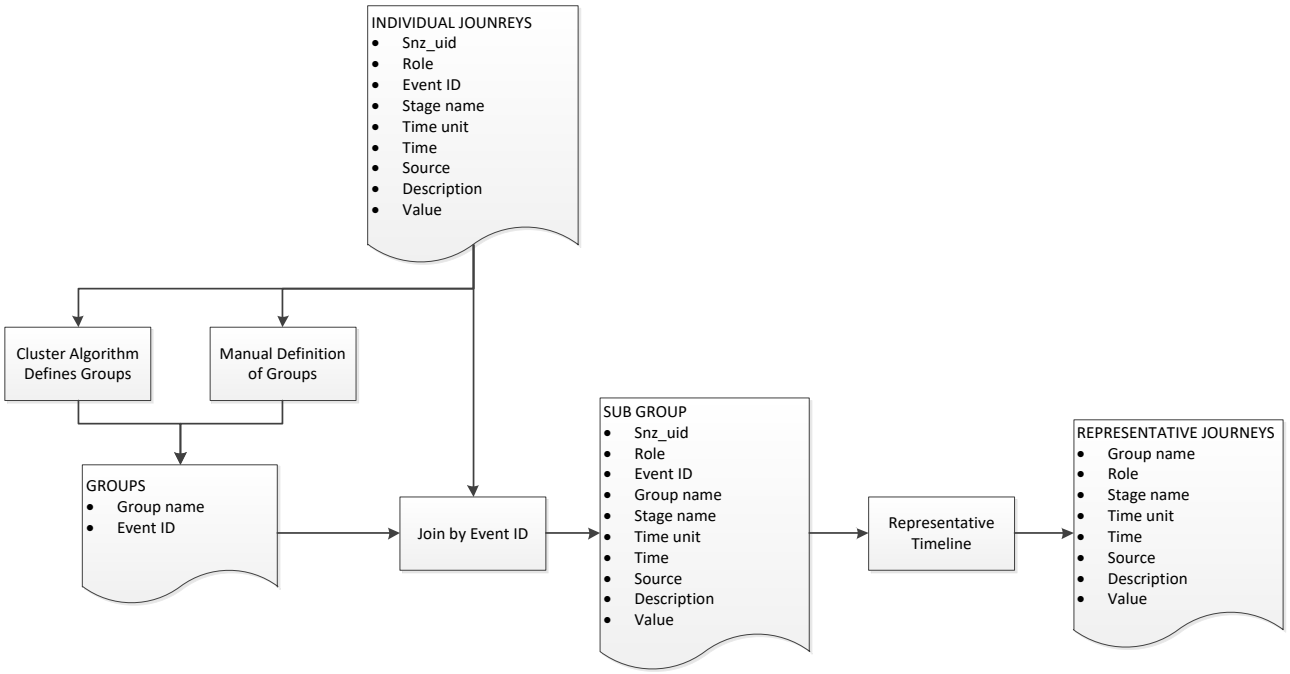
Where a representative timeline must be constructed, the method chosen for this construction will determine which features of the individual resolution timelines are preserved. Summarising each measure independently is more straightforward but is less likely to preserve interactions between different measures (for example: the waiting time between a doctor's visit and a follow up appointment with a specialist). Hence the importance of preserving interactions between specific measures needs to be identified early as it informs the choice of summary methodology.

At present, we do not have a summary methodology that combines individual timelines to produce a representative timeline in a way that preserves all the features of the original timelines. Hence the choice of summary methodology depends on the timeline features that are of most relevance to the research. For example: research into waiting times would likely prefer a method that ensures the gaps between events are representative even though it may result in less representative durations of events.

It will be rare for every individual resolution timeline to include observations of every measure. This impacts how measures are summarised. Rather than summarise a measure over all members of the group, we recommend reporting the proportion of timelines with each measure and the representative measure for those timelines that contain it. For example: if half of timelines contain two address changes and half contain no address changes, then it could be misleading to show a representative timeline with only a single address change. In this example, we would show a representative timeline with two address changes and report that half of the contributing timelines have address changes.

The following figure provides an overview of this process. A full description can be found in the appendix.

Figure 2: Producing representative timeline outputs



For our application to South Auckland, we began within individual level timelines within the secure environment of an IDI datalab. However, to respect the privacy of the people who appear in the data the construction of representative timelines also served to confidentialise the timelines prior to their release from the secure environment.

No technique for confidentialising timelines existed for IDI data prior to this work. In conversation with Stats NZ, SIA developed an approach for taking groups of individuals and for each part of the timeline determining the number and timing of events that best reflects the whole group and protects the privacy of the individuals who make up the group. This approach focuses on preserving the timing of events relative to the reference date of the birth.

Following similar ideas to algorithmic clustering, we defined a measure of similarity between timelines. This is then used to consider how well a representative timeline reflects the individual timelines in the group. Various candidate timelines were produced as part of an iterative process, and the variation that was most like the group of individual timelines is selected as the representative timeline. A mathematical description of this process is given in the appendix.

Timelines require interpretation to create insight

Our discussion above has focused on the preparation of information for creating analytical timelines. But construction of timelines is not enough by itself. How meaning is inferred from timelines requires consideration, as they can be more difficult to interpret than other, more traditional, statistics.

A timeline can contain a lot of information: The presence, absence, number, frequency, duration, prevalence, and type of events across different people and different timelines. Identifying the relevant details can be difficult when considering all this information at once. To make things easier, two general approaches merit consideration: Understanding a single experience in detail and making comparisons between experiences.

Understanding a single experience in detail helps the viewer to realize the many different components that contribute to the timeline. This is important where the timeline brings together data that was previously separate – such as combining different data sources, or from different family members. In this approach insights arise from viewers understanding how previously separate measures coincide.

Comparisons between experiences help the viewer to identify norms, as well as consistency and variation between experiences. An effective starting point for an investigation is an average or expected timeline. Even when the lived experience is already well understood, this helps establish a common point of comparison with insights arising from differences between timelines.

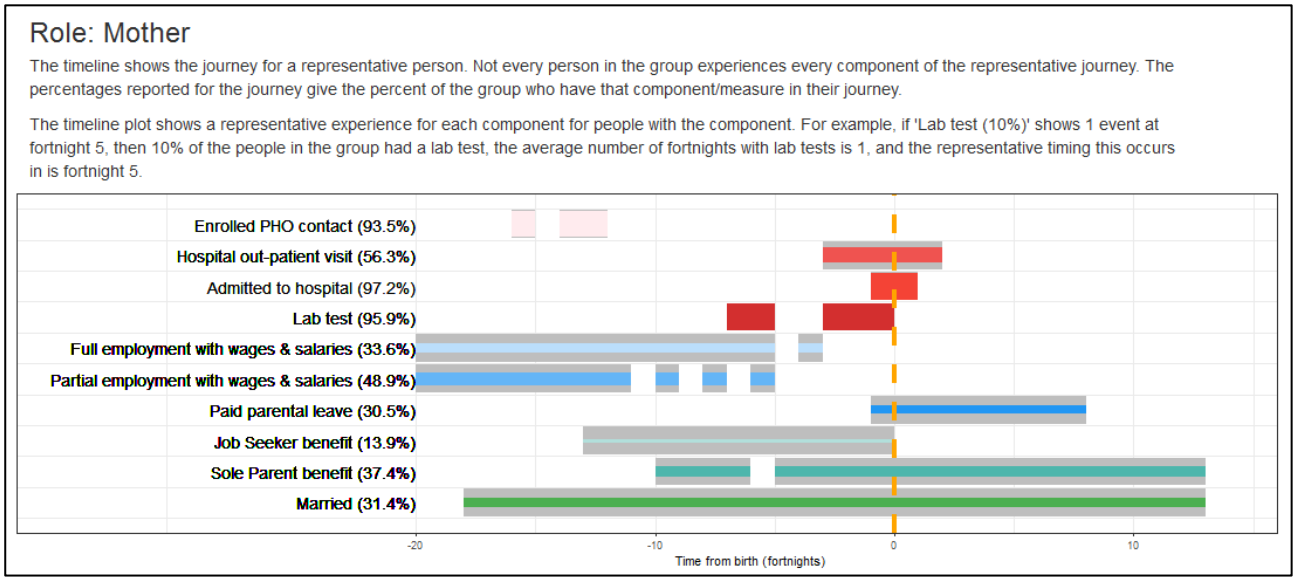
Visualisation of timelines supports interpretation

A visual representation of timelines is important as it provides the main way that people can examine and interpret the information. The exact choice of visualisation depends on the data to be displayed, the audience, and the insights that are being sought.

For our application to South Auckland, SIA built an interactive visualisation tool using the Shiny package in R. The tool was built to enable SIA and TSI staff to explore the data from different perspectives. It presented information in two forms: the first focused on a single experience (where all the roles that share a representative timeline can be viewed at once), and the second focused on comparison (where multiple representative timelines can be viewed at once but limited to a single role).

An example of the visualisation provided by the tool is given below. It is important to note that, like the timeline construction methodology, the visualisation tool is a proof of concept at the time of writing.

Figure 3: Example of timeline visualisation



In addition to the timeline visualisation displayed above, the tool includes a range of controls for the user to select the contents of the visualisation. This enables more targeted investigations as a timeline can be constructed containing only specific measures of interest, and the visualisation can be focused on specific groups or roles of interest (for example, viewing the timelines for a father and mother side-by-side, or making comparisons between timelines for mothers with different levels of education).

The visualisation tool is independent of the data preparation methodology. When the tool is run, it loads data from two Excel files: a timeline data file and a display configuration file. Any data that meets the format requirements of these Excel files can be loaded and visualised by the tool. Code for the visualisation tool is available on SIA's GitHub page[7] along with example data to guide the preparation of new inputs.

Interpreting timelines is best done with a range of expertise

As timelines describe a human experience, a human perspective is needed to identify what parts are significant. We have found the interaction of different types of expertise, contextual and technical, to be significant in interpreting timelines. Contextual expertise, understanding the people in the study population, is important for asking meaningful questions of the data and ensures that insights are applicable to the study population. Technical expertise, understanding the origin and manipulation of the data, is important for understanding the boundaries of the data and ensuring the appropriate caveats are applied.

For our application to South Auckland, the contextual expertise was provided by TSI and the technical expertise was provided by SIA. As they were connected to families and service providers, TSI's exploration of the timelines was guided by their knowledge of what was important to people in South Auckland. As we had conducted the analytics with the IDI, SIA tested any potential

[7] https://github.com/nz-social-investment-agency/timeline_visualisation

conclusion to ensure that all insights were consistent with the construction methodology. By working in partnership together our results were more robust than either organisation could have accomplished alone.

References

Abbott, Andrew, and John Forrest. (1986). Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History*, 16(3), 471-494.

Abbott, Andrew. 1995. "Sequence Analysis: New Methods for Old Ideas." *Annual Review of Sociology* 21:93-113.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150-3152.

Gabadinho, Alexis, Matthias Studer, Gilbert Ritschard, and Nicolas S. Müller. (2010). Sequence analysis for social scientists, Introduction to sequence analysis. Department of Econometrics, University of Geneva, Switzerland.

Giegerich, Robert, Stefan Kurtz, Enno Ohlebusch, Jens Stoye, Sven Rahmann, Eyla Willing, Peter Husemann, Roland Wittler, and Katharina Westerholt. (2012). *Sequence Analysis I + II, Lecture notes*. Faculty of Technology, Bielefeld University, Germany.

Mika, Sven, and Burkhard Rost. (2003). UniqueProt: creating representative protein sequence sets. *Nucleic acids research*, 31(13), 3789-3791.

Social Investment Agency. (2018a). *Are we making a difference in the lives of New Zealanders – how will we know? A wellbeing measurement approach for investing for social wellbeing in New Zealand*. Wellington, New Zealand.

Social Investment Agency. (2018b). *What you told us…findings of the 'Your voice, your data, your say' engagement on social wellbeing*. Wellington, New Zealand.

Stats NZ. (2016). *Microdata output guide (Fourth edition)*. Available from www.stats.govt.nz. ISBN 978-0-908350-68-1 (online)

Stats NZ. (2017). *Integrated Data Infrastructure: Overarching privacy impact assessment*. Retrieved from www.stats.govt.nz. ISBN 978-1-98-852818-2 (online)

The Southern Initiative, and The Auckland Co-Design Lab. (2017). *Early Years Challenge, Supporting Parents to give Tamariki a great start in life*. Auckland, New Zealand.

Appendix: Technical Details

Formal description of procedure for merging individual timelines to make representative timelines

Let A and B be ordered binary sequences (timelines) of equal length, $A = [a_1, a_2, \dots, a_n]$ and $B = [b_1, b_2, \dots, b_n]$ where $a_i, b_i \in \{0, 1\}$. We define the distance between timelines A and B as:

$$D(A, B) = \sum_{i=1}^{n} \min_j \{|i - j| : a_i = b_j\}$$

Let A denote a candidate representative timeline, B_k denote the k-th individual timeline and $A[i, j]$ denote the ordered sequence A with the values in the i-th and j-th positions swapped. Then a representative timeline is constructed according to the following greedy optimisation procedure:

1. Old timeline <- A
2. Old distance <- $\sum_k D(A, B_k)$
3. Repeat until stopped
    a. New timeline <- old timeline
    b. New distance <- old distance
    c. For($i \in 1{:}n$)
        i. If($a_i \neq a_{i+1}$)
            1. Candidate timeline <- $A[i, i+1]$
            2. Candidate distance <- $\sum_k D(A[i, i+1], B_k)$
            3. If(candidate distance < new distance)
                a. New timeline <- candidate timeline
                b. New distance <- candidate distance
    d. If(New distance = Old distance)
        i. STOP
    e. Otherwise
        i. Old timeline <- new timeline
        ii. Old distance <- new distance
4. Return old timeline as solution

List of measures used in our initial study of families in South Auckland

A range of measures are included in the timeline analysis. A collection of additional measures, not listed here, were used to provide supporting information.

Source	Measure	Description
SNZ	Address change	A notification for a different address reported to any government organisation, collated by SNZ
ACC	Accident	From claim submitted to ACC, dated by accident date
DIA	Married	Marriages recorded by DIA from the marriage date until the dissolved date (if any)

DIA	Civil Unions	Civil Unions recorded by DIA from the union date until the dissolved date (if any)
HNZ	Social housing application	All household members on a new social housing application, regardless of receipt or whether on other applications (Note: joined across all three system ID numbers)
HNZ	Living in a social house	Where social housing snapshots record a person in the same house two consecutive months, they are recorded as living at the address between these two dates
IRD	Employment	From Employer monthly schedule (EMS) returns for wages and salaries (W&S). The first of the month is used as the start date unless an employee start date is provided prior to (and within 60 days of) the return date. The end of the month is used as the end date unless an employee end date is provided after (and within 27 days of) the return date.
IRD	Partial employment	As per employment but where gross earnings are less than $2640 for the month (approx. minimum wage full time)
IRD	Full employment	As per employment but where gross earnings exceed $2640 for the month
IRD	Paid parental leave	As per employed but EMS labelled as paid parental leave instead of wages and salaries
MSD	Tier 1 benefit receipt	Main benefit receipt as per Marc de Boer's work incorporated in the SIAL. Dated from start to end date of receipt, where the recipient is the sole recipient, the primary recipient (with a partner) or the partner of the recipient
MSD	Tier 2 benefit receipt	Supplementary benefit reported to MSD from start to end date of receipt
MoE	Education enrolment	For targeted training, industry training, post-secondary and tertiary study, from the programme/placement start date until the end date
MoH	Contact with enrolled PHO	From PHO enrolment records, all distinct 'last consult' dates. Note multiple visits close together will not be observed
MoH	Contact with non-enrolled PHO	Inferred from general medical subsidies (GMS) records, all distinct visit dates
MoH	Active registration with a PHO	From PHO enrolment records dated from enrolment date to last consult date, for each practice a person has been enrolled in.
MoH	ED visit	From national non-admitted patient collection (NNPAC), dated by service date where the event type is labelled as ED and record is not marked 'did not attend'
MoH	Outpatient visit	As per ED visit but labelled as outpatient
MoH	Community visit	As per ED visit but labelled as community visit
MoH	Hospital admission	From publicly funded hospital discharges, labelled by start and end date
MoH	Lab test	From lab test claims, counts the number of distinct tests

		performed, then collapsed into an indicator for any tests in a given fortnight
MoH	Programme with an alcohol and drug team	A record in mental health care data (PRIMHD) with a team described as an 'alcohol and drug team'. Dated by activity start and end date. No controls for 'did not attend' or 'attended as family member of patient'.
MoH	Programme with a maternal mental health team	A record in mental health care data (PRIMHD) with a team described as a 'maternal mental health team'. Dated by activity start and end date. No controls for 'did not attend' or 'attended as family member of patient'.
MoH	Pharmacy dispensing – antidepressants	Dispensed pharmaceuticals, dated by dispensed date, where TG code for level 1 is 'nervous system' and for level 2 is 'antidepressants' Note the distinction between prescriptions and dispensing
MoH	Pharmacy dispensing – contraceptives	Dispensed pharmaceuticals, dated by dispensed date, where TG code for level 1 is 'genito-urinary' and for level 2 is 'contraceptives'
CYF	Report of concern	Reports of concert to CYF (intakes) that fall under section 15 for Reporting of ill-treatment or neglect of child of young person
Police NIA	Exposure to family violence	From 111 calls, where incident/offence code is one of: • common assault (domestic) • husband rapes wife • unlawful sexual connection with spouse • cruelty to child • domestic dispute This does not distinguish between offender and victim, and is known to be an under count
MoJ	Court hearing	Date of first court hearing, and last court hearing where different, for each charge Intermediate court hearing dates are not available
Corrections	Major management periods	The most serious sentence, from period start to end classified as one of: • community sentence • detained sentence • home detention sentence • under conditions • under supervision

Figure 4: SIA's method for constructing individual resolution timelines

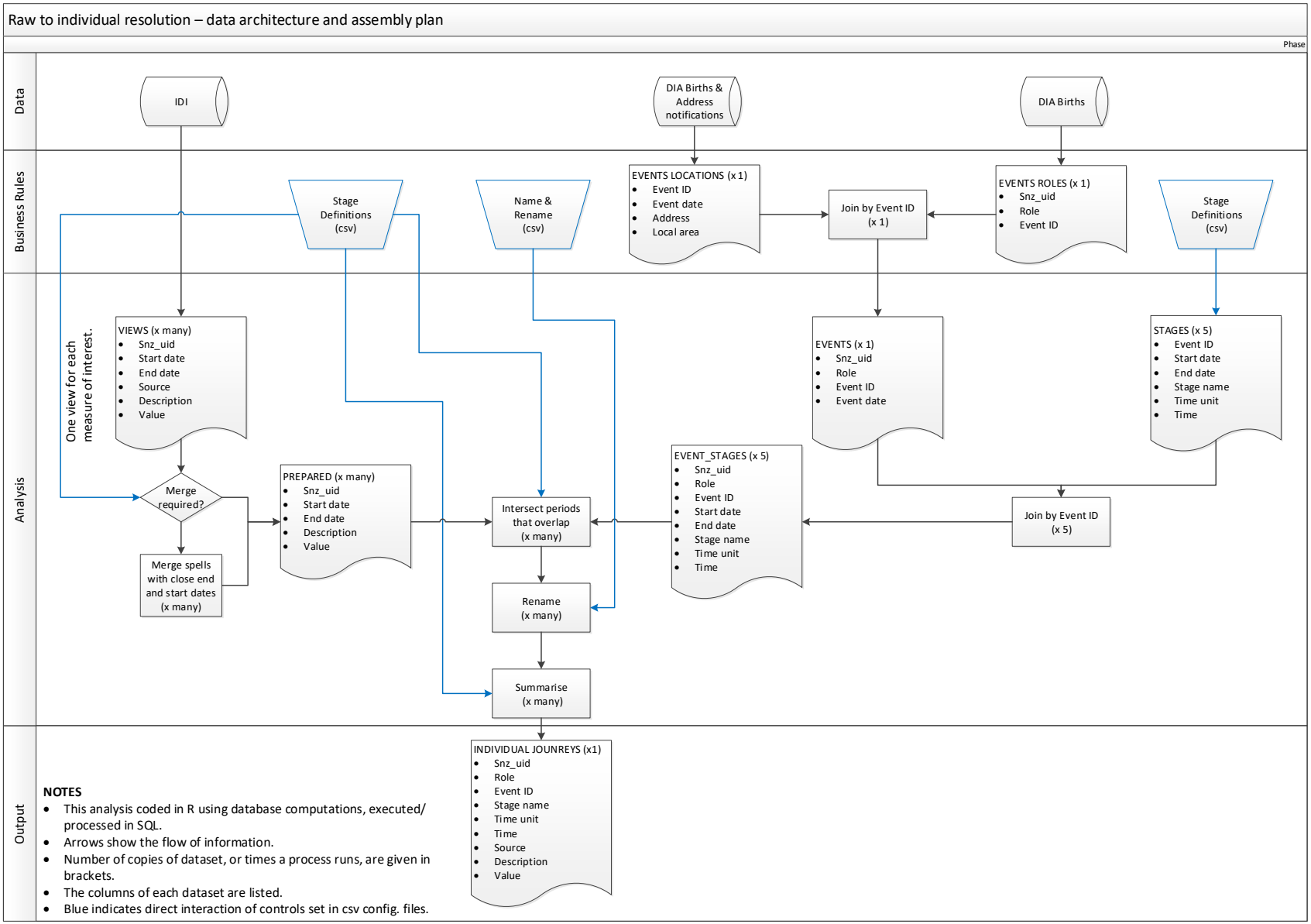


Figure 5: SIA's method for summarising to group timelines